DOCUMENT RESUME

ED 209 328 TM 810 870

AUTHOR Wheeler, Patricia; Harris, Abigail
TITLE Comparison of Male and Female Performance on the ATP Physics Test.
INSTITUTION College Entrance Examination Board, New York, N.Y.
REPORT NO CB-R-81-4
PUB DATE 81
NOTE 51p.
AVAILABLE FROM College Board Publications Orders, Box 2815, Princeton, NJ 08541 ($4.00).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Achievement Tests; College Entrance Examinations; Difficulty Level; *Item Analysis; *Performance Factors; *Physics; Science Tests; *Sex Differences
IDENTIFIERS *Physics Achievement Test (ATP)

ABSTRACT

This exploratory study on the College Board's Admissions Testing Program (ATP) Physics Test can be divided into two main parts, each designed to address a specific set of questions: Part I, Are there any systematic differences in male/female performance on individual items or subgroups of items that can help in interpreting the differences between males and females in overall performance on the test?, and Part II, Are there any factors that, when related to performance data on individual items or subgroups of items, can help in interpreting the overall performance differences and perhaps provide insights into unique characteristics of the female physics student? Questions from Part I are addressed by performing a standard test analysis following procedures associated with classical test theory. The second part of the investigation requires considering the available ATP data in the context of what is known from more general research of theories. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209328

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM SE

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

College Board Report

No. 81-4

Comparison of Male and Female Performance on the ATP Physics Test

Patricia Wheeler
Abigail Harris

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

P. K. Henkel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Comparison of Male and Female Performance on the ATP Physics Test

Patricia Wheeler

Abigail Harris

Educational Testing Service

College Board Report No. 81-4

College Entrance Examination Board, New York, 1981

We wish to express our appreciation to systems analysts Bernd Hasenkamp and Roger Ervin for their assistance in developing our data file; to science test development specialist George Pfeiffenberger for providing information on the test content; to research assistant Philip Makau for doing numerous calculations and preparing graphs and tables; to statistical consultants, Richard Zoller and John Bianchini; to the project's physics consultants Claire Ellen Max and Paul C. Wheeler, both physicists at Lawrence Livermore National Laboratory of the University of California; and to Nancy Castille for typing this report.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publication Orders, Box 2815, Princeton, New Jersey 08541. The price is $4.00

Copyright © 1981 by College Entrance Examination Board.
All rights reserved.
Printed in the United States of America.

CONTENTS

List of Figures iv

List of Tables v

Overview 1

Literature Review 3

- Participation in Science 4
- Performance in Science 4
- Psychometric Indicators of Performance Differences 6

Description of the Sample of Candidates 6

Description of the ATP Physics Test 7

Data Analysis and Results 11

- Data Preparation 11
- Data Analyses--Part I 11
- Results--Part I 11
- Data Analyses--Part II 25
- Results--Part II 26

Summary 34

References 39

LIST OF FIGURES

1 Mean Scaled Scores for Male and Female Candidates on the ATP Physics Test, 1965-66 through 1977-78 2

2 Candidate Volume Levels for the ATP Physics Achievement Test, 1965-66 through 1977-78 2

3 Delta Values on the ATP Physics Achievement Test for Male and Female Test Candidates, December 1978 16

4 Patterns of Test Speededness for Male and Female Candidates, ATP Physics Test, December 1978 20

5 Distribution of Subscores by Content Area for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates 27

6 Distribution of Subscores by Bloom Categories of Cognitive Complexity for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates 28

7 Distribution of Subscores by Item Format for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates 29

8 Mean Scores for Male and Female ATP Physics Test Candidates by Content Area by Number of Semesters of Physics 31

9 Mean Scores for Male and Female ATP Physics Test Candidates by Bloom Categories of Cognitive Complexity and on Total Test by Number of Semesters of Physics 32

10 Mean Scores for Male and Female ATP Physics Test Candidates by Item Format by Number of Semesters of Physics 33

11 Means of Total Formula Scores and Quantitative and Qualitative Subscores by Number of Years of Math at High School Graduation by Sex for Grade 12 ATP Physics Test Candidates 35

LIST OF TABLES

1	Level of Education	8
2	Number of Semesters of Physics Course Work	8
3	Number of Years of Math Course Work to Be Completed in High School	9
4	Ethnic Distribution of ATP Physics Candidates Who Completed Student Descriptive Questionnaire	9
5	Number of Items by Type of Item in ATP Physics Test	10
6	Summary Statistics by Sex for ATP Physics Test	12
7	Frequency Distribution of Deltas by Sex	12
8	Item Statistics for Males and Females for ATP Physics Test	14-15
9	Outliers for ATP Physics Test ($\|D_i\| > 0.63$)	17
10	Frequency Distribution of Biserial Correlations by Sex	18
11	Data on Speededness of the ATP Physics Test	19
12	Items Omitted by at Least Twice as Great a Proportion of Women as Men	20
13	Items Omitted by a Significantly Greater Percentage of Women than Men	23
14	Average Percent Omit for Male and for Female Candidates by Item Format	23
15	Percent Choosing Distractors for Items Where Males and Females Differ in Their Most Popular Distractor	24
16	Summary Statistics for Formula Scores by Achievement Band	26

OVERVIEW

Sizable performance and participation rate differences have existed between male and female candidates on the College Board Admissions Testing Program's (ATP) Physics Achievement Test for many years (Pfeiffenberger, 1974). As illustrated in Figure 1, the discrepancy in performance has increased gradually from a 33 point scaled score difference in the mid-1960s to almost twice that in recent years. The facts that the test is elective and that female candidates take this test in much smaller numbers than do male candidates (see Figure 2) make this discrepancy even more perplexing if one speculates that the women who do take the test represent a select group of women, presumably the ablest.

The underrepresentation of women evidenced on the ATP Physics Test is indicative of a pattern that shows up in higher education programs and in physics careers. Women receive only 5 percent of the doctorates awarded in physics (Andersen, 1976) and only one out of 20 or 30 physicists is a woman (Max, 1978). Within physics careers, women are also underrepresented in the more prestigious and lucrative jobs. A larger percentage of female than of male physicists are employed in two- and four-year colleges, while a larger percentage of male than of female physicists are employed by private profit-making companies (Centra, 1974).

The ATP Physics Test is a 75-item multiple-choice test that covers material typically taught in a one-year high school course. Topics covered include mechanics, electricity and magnetism, optics and waves, heat and kinetic theory, and modern physics. ATP Achievement Tests such as this one are used by colleges and universities as part of the selection and early placement process in academic programs. Also, in conjunction with other data, it is used by students and counselors as a guidance tool for gauging student potential. Students can use their test scores to select colleges and then apply only to those colleges where their performance is similar to that of enrolled students. The extent to which this self-selection is occurring is unknown, but there is evidence that high scoring candidates tend to apply to and be accepted by more prestigious institutions, whereas lower scoring candidates apply to and enroll in less prestigious colleges (Hartnett and Feldmesser, 1980).

If the self-selection phenomenon holds true in physics, it implies that the higher scoring candidates apply to and are accepted by colleges with higher quality physics programs and presumably are able to pursue more prestigious and lucrative careers after graduation than lower scoring candidates are able to do. Therefore, since women are proportionately underrepresented among the high scoring candidates, the observed discrepancies in physics test performance tend to perpetuate the differences in participation in physics in both higher education and careers. Investigating performance differences on the ATP Physics Test is a critical first step to understanding and possibly combating these patterns. The persisting differences in performance over a 13-year period demonstrate a need to investigate factors in the test that might contribute to or illuminate sources of such differences.

This exploratory study on the ATP Physics Test can be divided into two main parts, each designed to address a specific set of questions:

Part I. Are there any systematic differences in male/female performance on individual items or subgroups of items that can account for or help in interpreting the differences between males and females in overall performance on the test?

Part II. Are there any factors, such as number of semesters of physics or math, or overall level of performance on the ATP Physics Test that, when related to performance data on individual items or subgroups of items, can help in interpreting the overall performance differences and perhaps provide insights into unique characteristics of the female physics student?

Part I covers a set of several more discrete questions:

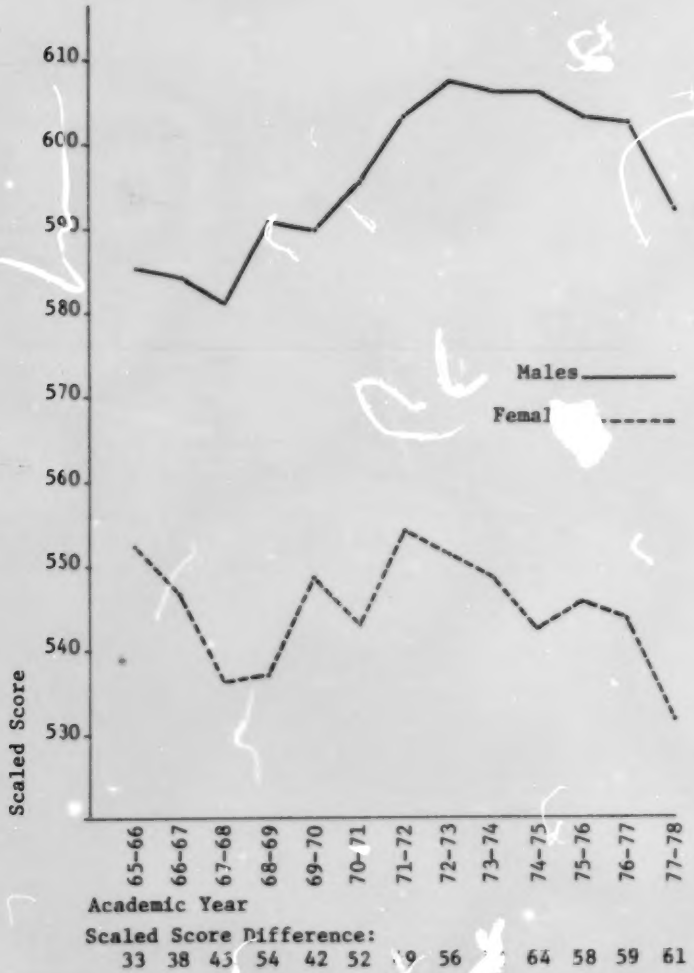

FIGURE 1. Mean Scaled Scores for Male and Female Candidates on the ATP Physics Test, 1965-66 through 1977-78

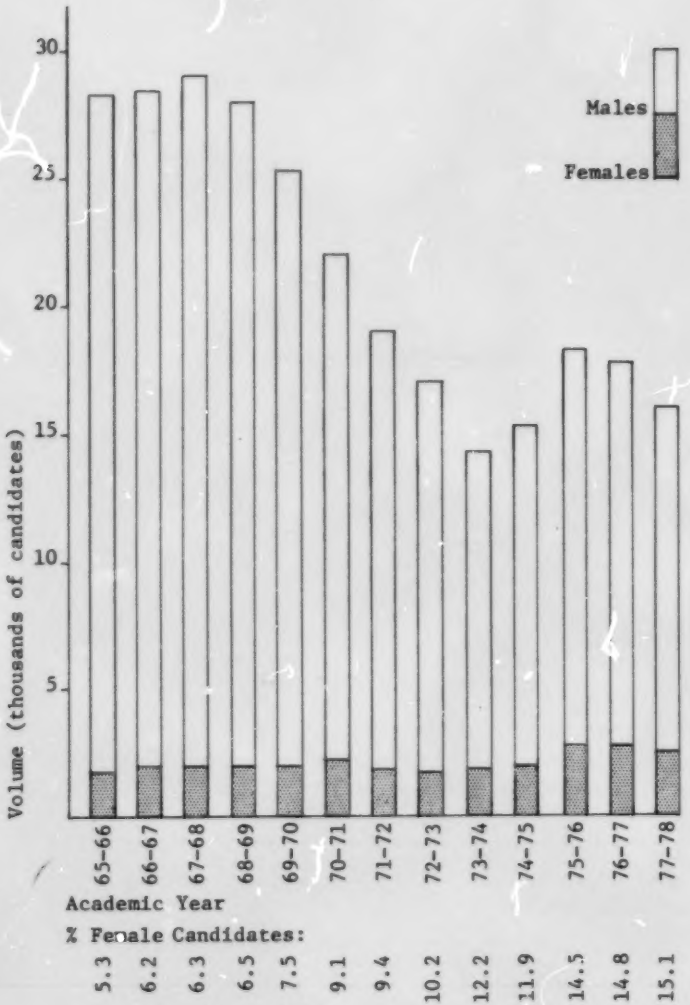

FIGURE 2. Candidate Volume Levels for the ATP Physics Achievement Test, 1965-66 through 1977-78

A. How do the test statistics (e.g., mean formula score, internal consistency reliability, standard error of measurement) compare for men and women?

B. Are there items that are differentially difficult for males and females? If so, are there any distinguishable characteristics associated with these items?

C. Are there items that discriminate differently for male and female candidates? If so, do these items have any discernable similarities?

D. Are there similar patterns of performance with regard to speededness and omits for men and women?

E. Does an error analysis reveal a relationship between sex and choice of distractor?

These questions can be addressed by performing a standard test analysis following procedures associated with classical test theory.

The second part of the investigation is less direct than the first. It requires considering the available ATP data in the context of what is known from more general research or theories. Because of the virtual absence of past research on sex differences in physics, this section is, by necessity, exploratory; one can only infer or hypothesize from research in related fields. For example, what happens when males and females with the same overall test performance are compared? Research shows that, in general, high school males outperform high school females in math (Fennema and Sherman, 1977; Maccoby and Jacklin, 1974). Can one expect, therefore, that a greater proportion of the items that the men answer correctly on the ATP Physics Test are quantitative items, whereas women do relatively better on the qualitative items? Are there content areas or item formats in which women perform as well as or better than men? Thus, the following questions are explored as part of this study's second major line of inquiry:

A. Do male and female candidates who achieve the same level of total test performance arrive at their scores in similar ways, e.g., do they have similar patterns of subtest score performance?

B. Does controlling for the number of physics courses or math courses reduce the male/female performance discrepancy on the total test or on subsections of the test?

Investigating these questions will clarify the strengths and weaknesses of the female candidates compared to the male candidates. It is anticipated that this exploratory study will lead to a greater understanding of the underrepresentation of women in the candidate pool and in the field of physics in general.

LITERATURE REVIEW

The existence of sex differences in physics participation and performance has been documented in previous studies. Several efforts have been undertaken to explain the relative absence of women in the sciences (Casserly, 1978; Keeves, 1973; Kelly, 1976; Max, 1978; Rossi, 1965; Sells, 1976; Stake and Granger, 1978). However, with one exception (Carlson, 1976), virtually none of these studies have sought to relate these participation differences to differences in test performance. Moreover, although sex differences in test performance have been explored in depth in other areas, especially in math and verbal ability, few studies have looked closely at performance differences in physics and other sciences. Studies looking at performance differences have employed different methodologies and considered different aspects of the test (e.g., content, speededness, difficulty level), making comparison of results from various studies more difficult.

Participation in Science

A variety of explanations have been offered to account for sex differences in high school participation in the sciences. The most frequently cited reasons link the participation discrepancy in the sciences to long established sex differences in math participation and performance (Maccoby and Jacklin, 1974). Investigators have drawn upon such theories as sex role socialization, sex role expectations, and modeling to explain sex differences in science and math attitudes, interests, and experiences (Casserly, 1978; Kelly, 1976). Differences in reinforcement are particularly important because, for the most part, science and math courses are sequenced. That is, one course builds on the skills and knowledge acquired in previous courses. If a student falters or if his/her involvement is disrupted at any point in this sequence, it is frequently difficult, if not impossible, for the student to pick up where he/she left off. This results in increasingly lower participation levels for both sexes, but particularly for women, as education level increases (Husen, Fägerlind, and Liljefors, 1974; Rossi, 1965; Sells, 1976).

Performance in Science

Although numerous studies have documented the existence of sex differences in performance in the sciences (Keeves, 1973; Kelly, 1978; National Assessment of Educational Progress, 1975) and in physics specifically (Carlson, 1976; Husen, Fägerlind, and Liljefors, 1974; Pfeiffenberger, 1974), very little has been done to investigate the sources of these differences. Some suggest that the differences in science and math may be linked and that they can be attributed to differences between the sexes in abilities; others suggest that the differences result from outside factors such as differences in academic background or out-of-school experience.

Proponents of an "ability" explanation for the sex differences point to those studies which have found sex differences in spatial visualization (Bock and Kolakowski, 1973; Maccoby and Jacklin, 1974; Stafford, 1961). That such differences might relate to differences in math achievement is illustrated by Donlon's (1973) analysis of the College Board's Scholastic Aptitude Test (SAT) scores. He found that sex differences in math aptitude scores (favoring males) could be cut in half if the questions on geometry were eliminated and only algebra knowledge questions were included on the test.

An earlier study (Walberg, 1969) linked verbal and spatial visualization abilities to the patterns of science achievement of high school students. Walberg reported data from the Harvard Physics Project in which high school girls scored higher than high school boys on tests of Scientific Processes and Understanding Science, while high school boys scored higher than high school girls on a test of Physics Achievement; the former tests involved mainly verbal skills, while the latter tapped visual-spatial skills.

Reports of two other major testing efforts include findings of sex differences in the sciences. One of the studies, which was conducted as part of the National Assessment of Educational Progress-NAEP (National Assessment of Educational Progress, 1975), investigated changes in science achievement between 1969 and 1973. Although scores declined steadily at three age levels over the four-year period, the gap between males and females remained constant at each level. The performance of 9 year old boys was, on the whole, 2-3 percent above that of girls. The average performance of 13 year old boys continued to be 4 percent above that of girls. At age 17, average male performance was 6 percent above that of the females. Unfortunately, no breakdown by sex by science content area was given in the 1975 NAEP report series. However, Carlson (1976) examined some of the item data associated with these differences. She reported that girls were scoring slightly above the national average on the biology items and well below the national average on the physics items.

The second major testing effort for which sex differences in science have been

reported involves data from the International Association for the Evaluation of Educational Achievement (IEA). Originally, Comber and Keeves (1973) reported on sex differences at three age levels on the total science and subtest scores for 19 countries participating in the IEA survey. They reported that boys performed consistently better than did girls and that the performance differences increased across the three age levels. Subtest differences were largest in physics and smallest in biology. Using Bloom's Taxonomy of Educational Objectives (1956), items on the IEA tests were classified into four behavioral categories: Functional Information, Understanding (Comprehension), Application, and Higher Processes. It was found that using these categories, sex differences were largest in the category labeled Understanding, with males outperforming females.

Husen, Fägerlind, and Liljefors (1974) focused on the Swedish data in the Comber and Keeves IEA study. They too found that the largest sex differences were in physics. In addition, they reported that these differences begin as early as grade three (with grade four girls performing lower than grade three boys) and increase as the students progress through school.

One of the conclusions of Comber and Keeves (1973) was that the weaker performance of girls could be due to girls doing less well than boys in Science Understanding. Their analysis did not separate the science subject areas by the Bloom categories with reference to sex differences. However, Carlson (1976) pursued this question in her dissertation work. She reexamined the item data on the IEA science tests for 10 and 14 year old boys and girls in 13 countries. Carlson found that the difference on Science Understanding items is more accurately described as girls doing less well on Physics Understanding items. The sexes performed similarly on Chemistry and Biology Understanding items. A second finding from her investigation was that girls performed less well than boys on science items dealing with electrical circuits, sound, magnetism, space, and the result of applying different forces.

In contrast to the above proponents of an "ability" explanation, proponents of an "experience" explanation for the sex differences in the sciences and in math often point to sex differences in activities that are believed to be conducive to the development of spatial visualization skills and math understanding (de Wolf, 1980; Fennema and Sherman, 1977). Studies of this type frequently relate the academic and nonacademic backgrounds of students to their resultant relative performance. Such studies are often confounded by the fact that students freely choose many of their own background activities.

Carlson (1976) used another method to investigate the "experience" explanation in the previously described study of sex differences on the IEA science tests. She examined the items showing significant sex differences as identified by a multivariate analysis technique (Hotelling's T^2). By applying a judgmental procedure, the significant items were categorized as biased or unbiased. Biased items were defined as those items where it could be argued that the opportunity to learn the content occurred outside of school. Items testing the differential out-of-school experiences of boys and girls were believed to reflect socialization differences and should be considered biased because they cannot be interpreted as measuring the inherent potential to learn science. Although this technique was useful in accounting for the sex differences in item performance for a few items, it did not account for the differences for the majority of items.

In summary, the research literature on performance in science suggests that tests which emphasize items requiring spatial visualization as opposed to items stressing verbal and algebraic content could mitigate against good test performance by girls (Donlon, 1973; Maccoby and Jacklin, 1974). It also appears that items emphasizing physics understanding, rather than physics knowledge or application, and items in some content areas of physics might account for a disproportionate amount of the sex differences in physics. Finally, there is some evidence that those items for which differences do occur should be examined to determine whether they reflect sex differences in out-of-school learning.

Psychometric Indicators of Performance Differences

One approach to investigating performance differences between groups on a test is to study the internal characteristics of the test by looking at item-statistics, such as p-values, deltas, and item-total correlations. More recently, item characteristic curves have been used in such studies (e.g., Lord, 1977; Scheuneman, 1975; Wright, Mead, and Draba, 1976). Other indicators such as completion rates (speededness), choice of distractors (error analysis), and omit patterns have been used in some studies (e.g., Wild and Durso, 1979). Of these procedures, the p-values and delta approaches have been used most often in empirical studies of performance differences between males and females.

The earliest studies of sex differences on item performance were experimental studies that employed few subjects and studied small numbers of items (e.g., Milton, 1959; Hoffman and Maier, 1966). In more recent studies, sex differences in some aspect of performance have been examined for large groups of test takers at the item level; items for which differences emerge are analyzed to see if some item content or format characteristic is common to these items. Most of the recent studies have been conducted on college-level admissions tests.

In studying the college-bound students of 1954, Cotfman (1961) examined items on the SAT-verbal aptitude test for gender bias and made predictions as to which items would favor men and which would favor women. The items were judged for traditional interest or skills of men and women. He determined the confidence intervals for the differences in percent pass (p) levels for each item, using a transformation formula, $\phi = 2 \arcsin p$, suggested by Walker and Lev (1954). Items that had differences in obtained transformed p-values outside the 5 percent level of confidence were identified and compared to the gender bias judgments. Seventeen of the 19 predictions based on these judgments were confirmed. Of the nine items involving mechanical knowledge, science, or business that were judged to be easier for males, eight were found to be easier for them. Of the 10 items involving personal feelings or personality characteristics that were judged as easier for women, nine were found to be easier for the female test takers

In a study of 1964 Scholastic Aptitude Test (SAT) takers, Donlon (1973) used difference in percent passing each item. Based on his analyses he found that verbal items favoring men tended to have scientific or practical affairs content, whereas verbal items favoring women tended to have human relations, humanities, or asthetical-philosophical content. Mathematics items favoring men tended to deal with "real world" topics; no math items favored women. These findings were supported in three more recent studies: for the 1974 SAT (Strassberg-Rosenberg and Donlon, 1975), for the 1978 SAT (Stern, 1978), and for the 1974 Graduate Record Examinations (Donlon, Hicks, and Wallmark, 1980). Although all of these studies used delta plots to examine performance differences, one of the studies (Stern, 1978) used an additional dimension to strengthen the methodology. In order to minimize the impact of large group differences in performance on delta values, she matched the groups using an external criterion, a verbal abili y test.

Wild and Durso (1979) examined the effects of testing time on verbal and quantitative scores of males and females on the Aptitude Test of the Graduate Record Examinations Board. The conclusions, based on the candidates tested at about 550 test centers in the United States in December 1976, were that increasing test time did not affect performance differences between males and females; however, the reduction in speededness from increased testing time did result in small score gains in both groups.

DESCRIPTION OF THE SAMPLE OF CANDIDATES

For this study, the data files of the December 1978 ATP Physics Achievement Test candidates were used. This date was selected because typically the December testing has a relatively large number of candidates compared to other test dates in any given year and the mean test performance is similar to that of the full year.

A total of 4,725 candidates took the ATP Physics Achievement Test on the December 1978 administration. Of these, 4,015 (85 percent) were males and 710 (15 percent) were females. As seen in Table 1, most of the candidates (94.3 percent) were high school seniors. Many of the candidates had completed only one semester of physics; very few had had more than three semesters of course work in physics (see Table 2). Female candidates had completed an average of 1.93 semesters while males had completed an average of 1.97 semesters of course work. This difference is significant, based on a chi-square analysis of the frequencies in Table 2, excluding "Other/Missing" ($\chi^2 = 7.84$; $p \leq .05$). It appears that the male candidates are more apt to take a third semester of physics, while females are more apt to stop after two semesters. However, no information is available on the types of physics courses they took.

The ATP Student Descriptive Questionnaire (SDQ) is an optional survey that provides additional data on demographic, background, and other candidate factors. About 90 percent of ATP candidates complete this survey at least once. Although this investigation was not designed originally to include SDQ data, two items on the SDQ were of particular interest in describing the sample of physics candidates: the number of semesters of math courses (since differences here could affect performance differences on a physics test) and ethnicity of the candidates. About 30 percent of the physics test candidates from the December 1978 administration completed all or part of the SDQ as part of their registration. Other candidates presumably completed the SDQ at an earlier administration or elected not to complete it. The results, based on the 30 percent sample, for math courses and ethnicity are presented in Tables 3 and 4, respectively.

Table 3 provides the number of males and females at each grade level who reported the number of years of math course work they had completed or expected to complete while in high school. As is evident, about 90 percent of both the males and the females were seniors who expected to complete four or more years of math by high school graduation.

The numbers of SDQ respondents in each ethnic group are given in Table 4. The proportions of males and females in each group are comparable. The majority of the respondents are white (about 75 percent), 12 percent described themselves as Oriental or Asian American, and 6 percent as black, Afro-American, or Negro.

DESCRIPTION OF THE ATP PHYSICS TEST

The ATP Physics Achievement Test is a 75-item multiple-choice test designed to cover the materials normally taught in a one-year high school physics course (College Entrance Examination Board, 1978). The test was designed for use as part of the selection process for college admissions and for placement in introductory college physics courses. Each item on the test has five choice selections. The candidate is given one hour to complete the test. The test is formula scored with a correction for guessing; the formula score equals the number right less one-quarter the number wrong. Formula scores are then converted to scaled scores and reported to candidates and the colleges they designate.

Items on the test were first classified by the test development staff at the Educational Testing Service office in Princeton, New Jersey. For this project, the classifications were reviewed and verified by two physics consultants prior to any analyses being run for this study. Several classification schemes were used. They are described here and listed in Table 5.

The test covers five major content areas: Mechanics, Electricity and Magnetism, Optics and Waves, Heat and Kinetic Theory, and Modern Physics. Mechanics includes items on kinematics, dynamics and statics, energy and momentum, forces and acceleration, circular motion and rotation, and vibrations and simple harmonic motion. Electricity and Magnetism covers electrostatics, circuits, and electromagnetism. Topics such as lenses and light and sound waves are included under Optics and Waves. Items on thermal properties and on

TABLE 1. Level of Education
of ATP Physics Test Candidates, December 1978

Level of Education	Males N	Males %	Females N	Females %	Total N	Total %
Grade 10	2	0.0	-	-	2	0.0
Grade 11	45	1.1	2	0.3	47	1.0
Grade 12	3,778	94.1	679	95.6	4,457	94.3
First Year College	47	1.2	8	1.1	55	1.2
Second Year College	24	0.6	4	0.6	28	0.6
Other	119	3.0	17	2.4	136	2.9
Total	4,015	100.0	710	100.0	4,725	100.0

TABLE 2. Number of Semesters of Physics Course Work
of ATP Physics Test Candidates, December 1978

Number of Semesters	Males N	Males %*	Females N	Females %*	Total N	Total %*
One	1,143	32.6	210	31.8	1,353	32.5
Two	1,462	41.7	310	46.9	1,772	42.5
Three	763	21.8	119	18.0	882	21.2
Four	139	4.0	22	3.3	161	3.9
Other/Missing	508	—	49	—	557	—
Total	4,015	100.1	710	100.0	4,725	100.1

* Excluding Other/Missing
χ^2 = 7.84, significant at $p \leq .05$

TABLE 3. Number of Years of Math Course Work to Be Completed in High School, ATP Physics Test Candidates, December 1978 (N = 1,199 Males; N = 262 Females)

Grade/Number Years of Math to Be Completed in High School	Males N	Males %	Females N	Females %
Grade 10:				
over 4 years	1	0.1	-	-
Grade 11:				
3 years	4	0.3	-	-
4 years	8	0.7	-	-
over 4 years	8	0.7	1	0.4
Grade 12:				
1 year	3	0.3	2	0.8
2 years	9	0.8	4	1.5
3 years	52	4.4	15	5.7
4 years	652	54.4	135	51.5
over 4 years	442	36.9	101	38.5
First Year of College:				
3 years	1	0.1	-	-
4 years	5	0.4	2	0.8
over 4 years	9	0.8	2	0.8
Second Year of College:				
4 years	2	0.2	-	-
over 4 years	3	0.3	-	-
Total	1,199	100.4	262	100.?

TABLE 4. Ethnic Distribution of ATP Physics Candidates Who Completed Student Descriptive Questionnaire, December 1978 (Data Available for 1,441 of 4,725 Candidates)

Ethnic Group	Males N	Males %	Females N	Females %
American Indian	3	0.3	-	-
Black	64	5.4	22	8.7
Mexican-Chicano	5	0.4	3	1.2
Oriental-Asian	134	11.3	32	12.6
Puerto Rican	9	0.8	3	1.2
White	892	75.1	179	70.5
Other	80	6.7	15	5.9
Total	1,187	100.0	254	100.1

TABLE 5. Number of Items by Type of Item in ATP Physics Test (Total = 75 Items)

Item Type	Number of Items
Content Areas:	
Mechanics	29
Electricity and Magnetism	16
Optics and Waves	16
Heat and Kinetic Theory	7
Modern Physics	6
Miscellaneous	1
Bloom's Taxonomy:	
Knowledge	25
Comprehension	17
Application	49
Format:	
Graphic	8
Numerical	17
Pictorial	24
Symbolic	3
Verbal	27
Qualitative vs. Quantitative:	
Qualitative	37
Quantitative	38

gases and kinetic theory are under Heat and Kinetic Theory. Modern Physics covers such areas as atomic and nuclear physics, photon, quantum phenomena, Compton effect, photoelectric effect, and relativity. The distribution of the number of items in each of these five content areas for the December 1978 test is given in Table 5. This form of the test contained one miscellaneous item that is of a general nature and overlaps several areas of physics.

Items were also classified according to Bloom's taxonomy into three levels: Knowledge (functional information), Comprehension (understanding), and Application. No items on this test were classified into Bloom's categories of higher level objectives (e.g., Synthesis, Analysis, Evaluation). Classification of items was more subjective in this case than with the content classifications. For example, many items could be answered through memory (Knowledge) if the student had been taught to memorize. Other students could have a greater understanding of the item beyond a simple repetition of the concept (Comprehension). Items that fit in two classifications were included in both classifications.

Item format was the third breakdown of items used for this study. Items were classified into five formats: Graphic, Numerical, Pictorial, Symbolic, and Verbal. Some items were counted under two formats when neither format dominated the other. Because of the low number of Symbolic items, separate analyses for this group of items were not performed. Items were also classified as either Qualitative or Quantitative.

DATA ANALYSIS AND RESULTS

Data Preparation

In order to answer the questions posed earlier, the records of all December 1978 ATP Physics Test candidates were extracted from the full ATP data file for that date. Two data sets were prepared. The first set contained item data for all female candidates and for a similar size sample of randomly selected male candidates. Regrettably financial constraints precluded the preparation of samples of males and females that were matched on an external criterion such as Scholastic Aptitude Test scores or high school rank. The mean formula score for this subsample of males is slightly, but not significantly, lower than that of the full male sample (38.19 vs. 38.66). This data set was used to respond to the questions in the first part of this study. The second data set, corresponding to Part II of this study, contained the records of all 4,725 candidates who took the ATP Physics Test in December 1978.

Data Analyses--Part I

Standard item analyses, adjusting the p-values, percent omitting, and delta values for items not reached, were run separately by sex in order to determine if there were any systematic differences in male/female performance on individual items or subgroups of items. The item analysis for females included the data for all 710 female candidates. Because of the high ratio of male to female candidates, the item analysis for males was performed on a random sample of one-fifth of the male candidates. This sample size (N = 802) is closer to that of the females and facilitates direct comparisons of the information obtained from the two item analyses. The findings, based upon these analyses, follow for each question in Part I.

Results--Part I

I-A. How do the test statistics compare for men and women? Summary statistics for male and female candidates are provided in Table 6. Both the mean and median reflect about a 10-point (two-thirds of a standard deviation) difference in performance, based on formula scores, between the male and female candidates.

The distribution is more skewed for women than for men, with the women's scores tending to pile up at the lower end of the distribution. The test appears to be highly reliable for both men and women. Male and female candidates have about the same standard error of measurement.

The r-biserial value for each item is a measure of that item's power to discriminate among different levels of performance, based on the total test score. The mean r-biserial value for all items for women is slightly lower than that for men.

Summary information on the item delta distributions for men and women is provided in Tables 6 and 7. Delta is a standard score index of item difficulty. For each item, the proportion of the sample answering the item correctly (p-value) is transformed, by reference to a table of the normal curve, to normal deviates (z) and from the normal deviates to delta values by the linear transformation, $\Delta = 4z + 13$ (Angoff and Ford, 1973). The easier an item is for a particular sample, the larger the item p-value and the smaller the corresponding item delta value. If one group's overall performance on a test is higher than another group's, then the mean of the item delta values for the higher scoring group will be smaller than the mean of the item delta values for the lower scoring group, as is the case in this instance. Of greater interest are group comparisons of the relative difficulty level of individual test items.

TABLE 6. Summary Statistics by Sex for ATP Physics Test, December 1978
(Based on Rights Minus 1/4 Wrongs)

	Male	Female
Number of Candidates	802	710
Mean Formula Score	38.19	27.99
Standard Deviation	15.94	14.34
Median Formula Score	37.05	25.98
Range*	-4 to 75	-1 to 73
Skewness	0.09	0.45
Reliability	.93	.9
Standard Error of Measurement	4.37	4.41
r-biserial:		
Mean	.52	.47
Standard Deviation	.10	.11
Delta:		
Mean	12.01	13.36
Standard Deviation	2.01	2.01
Percent Omit:		
Mean	11.24	18.65
Standard Deviation	7.98	12.06

*Possible range is -19 to 75

TABLE 7. Frequency Distribution of Deltas by Sex,
ATP Physics Test, December 1978
(N = 802 Males; N = 710 Females)

Delta	Males	Females
18.0 - 18.9	-	-
17.0 - 17.9	-	3
16.0 - 16.9	-	5
15.0 - 15.9	4	11
14.0 - 14.9	8	14
13.0 - 13.9	14	12
12.0 - 12.9	15	12
11.0 - 11.9	13	10
10.0 - 10.9	7	3
9.0 - 9.9	9	2
8.0 - 8.9	1	3
7.0 - 7.9	4	-
$\bar{X}$	12.01	13.36
s.d.	2.01	2.01

The percent omit is the proportion of candidates who reached an item, but did not select an answer. The mean percent omit for women is higher than for men. The differences in deltas, r-biserials, and percent omit values are discussed in the next three sections.

I-B. Are there items that are differentially difficult for males and females? If so, are there any distinguishable characteristics associated with these items? The item statistics obtained as part of the item analysis are provided in Table 8. Using a method described by Angoff and Ford (1973) to identify outliers (i.e., extreme items which are differentially difficult relative to the other items for male and for female candidates), a scatterplot of delta values was prepared (see Figure 3). Each point on this figure represents the delta values for males and females on one item. If the item was of equal difficulty for male and female candidates the point would fall on the 45-degree line (the solid line in Figure 3). Points above this line are harder for female candidates and points below harder for male candidates. Only one item falls below the 45-degree line (i.e., is easier for females than for males), two on the line (same delta for both females and males), and the remaining 72 above this line (easier for males). Female candidates find all but three items, from the easiest to the hardest items, more difficult than do the males.

The plot correlation is a measure of the degree of similarity of the delta ranks for male and female candidates. If two samples are not significantly different, the scatterplot will appear as an elongated ellipse and the correlation between the two sets of delta values will approach 1.0 (Angoff and Ford, 1973). The dashed line in Figure 3 is the major axis line of the ellipse formed by the plot of the 75 points; this line almost parallels the 45-degree line. The correlation for the points in Figure 3 is .94, indicating that the rankings for male and female candidates are very similar.

From Figure 3 it was difficult to identify visually any outliers. To identify more accurately those items that were relatively more difficult for male candidates and those that were relatively more difficult for female candidates, a series of calculations were made. The first step was to measure the perpendicular distance, D_i, from the major axis line ($y = 0.9473\ x + 2.0207$) of each point in the plot. The equation for D_i is:

$$D_i = \frac{Ax_i - y_i + B}{\sqrt{A^2 + 1}}$$

where $A = 0.9473$
$B = 2.0207$

The D_i values for each item are given in Table 8. Items with points below the major axis line (i.e., those that have positive D_i values) are those items that are relatively easier for the female candidates; those points above the major axis line (i.e., those that have negative D_i values) represent items relatively easier for males.

For this study, outliers were defined as those items that are at or beyond ± 1.25 standard deviations of the D_i distribution. The mean of the D_i distribution in this study was 0.00 (which it should be by definition) and the standard deviation was 0.5085. Therefore, outliers were defined as those items that have an absolute D_i value equal to or greater than 1.25 x 0.5085, or 0.6356. Fourteen items were identified as outliers on this basis. They are listed in Table 9.

The most interesting thing to note in Table 9 is the difference in item formats relatively easier for males compared to females. Based on past research on sex differences in quantitative ability (Aiken, 1976; Anastasi, 1958; Maccoby and Jacklin, 1974), one might expect that the items relatively easier for males would be the Numerical, Symbolic, and Graphic items, whereas those relatively easier for females would be Verbal and Pictorial items. The opposite was found in this study.

This finding is probably due to the fact that the female candidates electing to take

TABLE 8. Item Statistics for Males and Females for ATP Physics Test, December 1978 (N = 802 Males; N = 710 Females)

Item No.	Percent Pass		Percent Omit		Delta Values		D_i	r-biserials	
	Males	Females	Males	Females	Males	Females		Males	Females
1	90.3	86.2	2.1	3.8	7.9	8.7	.5839	.39	.33
2	81.7	60.4	3.7	10.2	9.3	12.0	- .8489	.60	.55
3	85.6	68.1	2.2	7.6	8.7	11.1	- .6082	.47	.42
4	72.8	58.5	1.7	5.7	10.6	12.2	- .1001	.43	.35
5	81.9	66.0	2.2	7.6	9.3	11.4	- .4134	.66	.63
6	67.0	51.3	4.8	11.0	11.2	12.9	- .1957	.60	.54
7	47.3	35.5	3.9	9.7	13.3	14.4	.1596	.54	.46
8	92.4	85.5	2.9	6.3	7.4	8.9	.0949	.34	.32
9	62.2	54.0	4.2	12.2	11.8	12.6	.4347	.39	.38
10	62.2	46.4	3.9	12.3	11.8	13.4	- .1460	.47	.51
11	56.1	36.1	7.2	17.9	12.4	14.4	- .4594	.53	.50
12	64.3	53.3	8.8	15.9	11.6	12.7	.2246	.52	.52
13	54.9	44.5	14.0	23.9	12.5	13.6	.1902	.63	.54
14	47.5	38.7	7.2	12.3	13.2	14.1	.3086	.28	.29
15	93.2	88.5	2.0	2.5	7.1	8.1	.4694	.50	.38
16	40.7	25.9	3.4	3.8	13.9	15.1	- .2990	.64	.50
17	84.5	65.9	3.1	8.0	9.0	11.4	- .6197	.52	.44
18	64.2	44.8	5.7	12.7	11.6	13.5	- .3562	.32	.34
19	70.5	78.0	10.3	11.0	9.8	9.9	1.0194	.62	.54
20	77.3	64.9	1.9	3.6	10.0	11.5	- .0046	.42	.54
21	76.1	62.4	2.5	5.9	10.2	11.8	- .0848	.35	.29
22	77.9	69.0	4.5	8.5	9.9	11.0	.2896	.46	.38
23	62.6	51.0	7.1	11.6	11.7	12.9	.1482	.51	.56
24	68.8	58.0	7.1	11.2	11.0	12.2	.1750	.52	.45
25	90.1	79.7	2.4	5.5	7.9	9.6	- .0694	.39	.43
26	65.0	55.4	8.6	9.7	11.5	12.5	.3010	.71	.61
27	56.0	41.3	16.5	24.5	12.4	13.9	- .0964	.66	.58
28	60.5	49.7	16.5	25.7	12.0	13.0	.2819	.49	.39
29	71.4	66.6	3.0	2.2	10.8	11.2	.7634	.52	.45
30	50.3	42.0	2.7	2.7	13.0	13.8	.3888	.55	.48
31	53.4	34.5	12.1	22.2	12.7	14.5	- .3257	.48	.30
32	81.7	63.4	8.2	20.0	9.4	11.7	- .5624	.58	.40
33	53.0	34.3	6.7	12.4	12.7	14.7	- .4708	.48	.41
34	46.1	38.3	12.9	18.2	13.4	14.2	.3735	.35	.45
35	79.1	74.7	11.2	13.3	9.8	10.3	.7290	.65	.54
36	68.7	61.4	1.9	2.9	11.0	11.9	.3927	.34	.41
37	80.2	62.2	6.6	16.2	9.6	11.8	- .4974	.44	.32
38	83.0	76.6	3.9	5.6	9.2	10.0	.5342	.62	.55
39	70.3	53.7	8.7	14.1	10.9	12.6	- .1842	.55	.45
40	51.9	27.8	3.2	7.4	12.8	15.3	- .8376	.49	.30

(TABLE 8, continued)

Item No.	Percent Pass		Percent Omit		Delta Values		D_i	r-biserials	
	Males	Females	Males	Females	Males	Females		Males	Females
41	66.9	37.9	10.7	32.9	11.3	14.2	-1.0706	.42	.51
42	56.9	30.3	31.4	57.8	12.3	15.1	-1.0363	.53	.36
43	54.8	29.2	16.9	34.5	12.5	15.2	- .9714	.56	.49
44	57.0	44.3	7.7	16.2	12.3	13.6	.0526	.52	.52
45	64.5	60.3	17.9	24.1	11.5	12.0	.6640	.56	.61
46	77.0	69.7	5.0	10.8	10.1	10.9	.4998	.52	.57
47	56.9	45.0	14.4	23.4	12.3	13.5	.1252	.53	.35
48	60.7	47.4	11.6	22.9	11.9	13.3	- .0047	.42	.48
49	61.4	58.7	6.3	10.6	11.9	12.1	.8665	56	.57
50	38.6	38.4	19.8	27.1	14.1	14.2	.8549	.60	.54
51	29.5	25.0	27.3	39.1	15.1	15.7	.4537	.53	.43
52	74.0	49.0	10.3	26.5	10.4	13.1	- .8910	.54	.37
53	37.0	19.1	13.4	17.2	14.4	16.5	- .6085	.44	.35
54	44.1	30.9	32.7	50.3	13.7	15.0	- .0009	.62	.61
55	52.4	43.4	13.1	21.6	12.9	13.7	.3926	.64	.64
56	41.6	28.0	18.2	28.0	13.9	15.4	- .1538	.63	.52
57	44.9	34.0	21.0	31.9	13.6	14.7	.1481	.52	.56
58	46.5	25.5	20.7	34.8	13.5	15.6	- .5740	.67	.70
59	46.6	36.2	10.6	19.8	13.4	14.5	.1557	.45	.50
60	42.7	32.4	22.5	33.9	13.8	14.9	.1404	.47	.50
61	39.8	21.9	18.8	29.5	14.1	16.1	- .5244	.65	.60
62	60.0	54.3	13.4	17.1	12.1	12.7	.5684	.52	.48
63	40.9	21.9	14.4	28.6	14.0	16.2	- .6658	.53	.39
64	35.9	18.3	24.3	39.6	14.6	16.7	- .6161	.64	.51
65	52.8	39.8	17.1	30.2	12.8	14.1	.0335	.57	.57
66	53.7	37.1	23.5	36.2	12.8	14.4	- .1843	.66	.62
67	45.7	35.3	12.5	15.7	13.6	14.6	.2207	.60	.60
68	39.1	26.7	21.4	38.0	14.3	15.5	.0487	.51	.32
69	25.2	15.5	2.6	3.7	15.9	17.2	- .0851	.69	.59
70	37.4	22.4	29.2	39.2	14.7	16.4	- .3296	.54	.43
71	44.9	28.4	16.1	29.1	13.9	15.6	- .2990	.51	.53
72	46.7	32.8	17.6	22.9	13.7	15.1	- .0735	.48	.42
73	40.2	44.3	22.0	22.3	14.5	13.9	1.3478	.30	.19
74	32.1	26.8	15.1	24.1	15.5	16.0	.5110	.46	.36
75	32.0	22.2	21.7*	21.3*	15.3	16.1	.3008	.45	.67

*Calculation based on assumption that all who reached item 74 reached item 75.

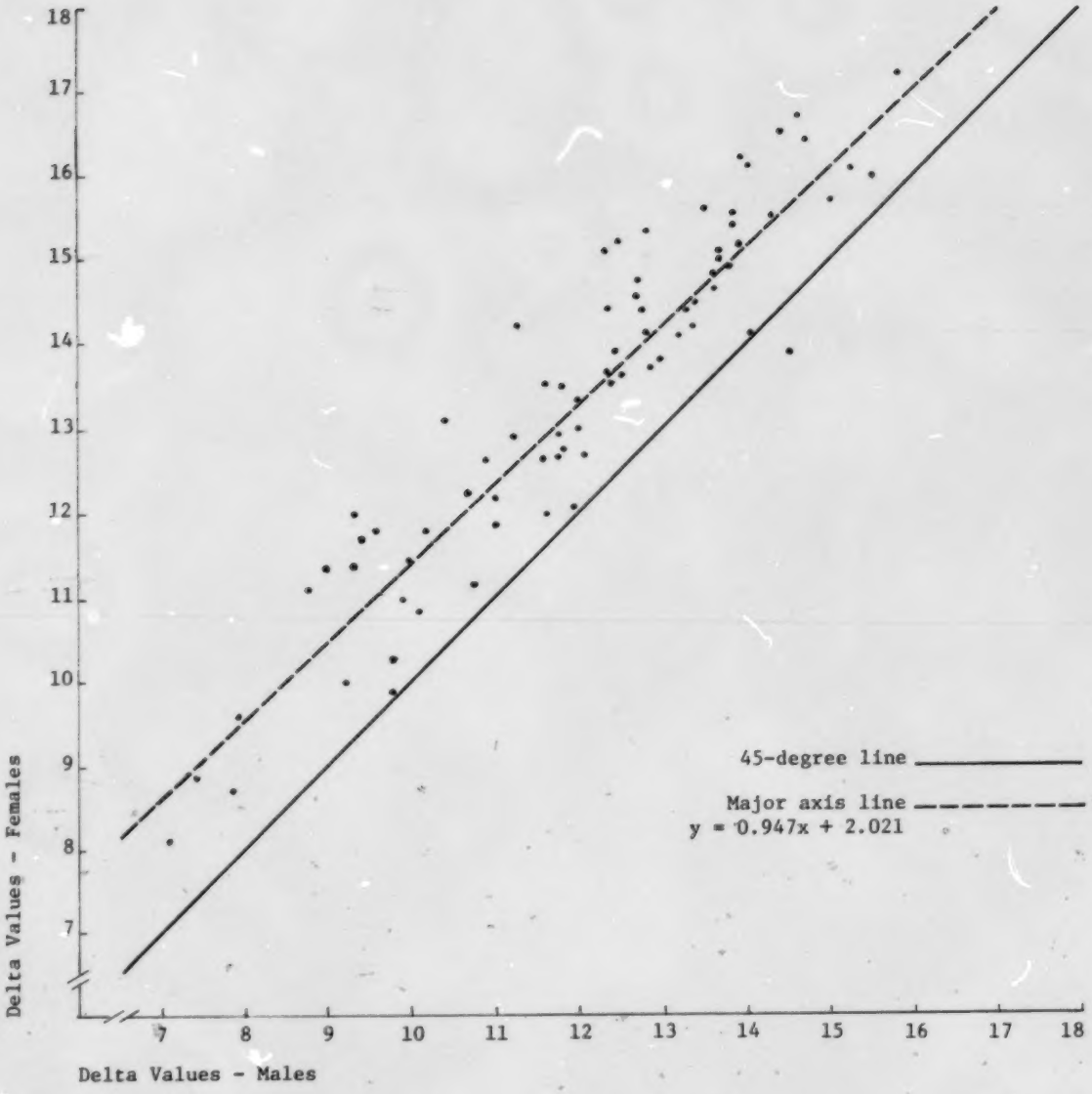

FIGURE 3. Delta Values on the ATP Physics Achievement Test for Male and Female Test Candidates, December 1978
(Number of males = 802; Number of females = 710)

TABLE 9. Outliers for the ATP Physics Test ($|D_i| > 0.63$), December 1978 (N = 802 Males, N = 710 Females)

	Item Content	Bloom Taxonomy	Item Format	Male Delta Value	Female Delta Value	D_i
Items Relatively Easier for Males						
2	Mechanics	Application	Pictorial	9.3	12.0	- 0.85
40	Electricity and Magnetism	Application	Pictorial	12.8	15.3	- 0.84
41	Electricity and Magnetism	Knowledge	Verbal	11.3	14.2	- 1.07
42	Electricity and Magnetism	Application	Pictorial	12.3	15.1	- 1.04
43	Modern Physics	Knowledge/ Comprehension/ Application	Verbal	12.5	15.2	- 0.97
52	Mechanics	Application	Numerical	10.4	13.1	- 0.89
63	Heat and Kinetic Theory	Knowledge/ Application	Verbal	14.0	16.2	- 0.67
Items Relatively Easier for Females						
19	Mechanics	Comprehension/ Application	Numerical	9.8	9.9	1.02
29	Mechanics	Application	Graphic/ Numerical	10.8	11.2	0.76
35	Other	Comprehension	Graphic	9.8	10.3	0.73
45	Heat and Kinetic Theory	Application	Numerical	11.5	12.0	0.66
49	Mechanics	Comprehension/ Application	Symbolic	11.9	12.1	0.87
50	Mechanics	Comprehension/ Application	Symbolic	14.1	14.2	0.85
73	Heat and Kinetic Theory	Application	Numerical	14.5	13.9	1.35

TABLE 10. Frequency Distribution of Biserial Correlations by Sex, ATP Physics Test, December 1978 (N = 75 items)

r-biserial	Males (N=802)	Females (N=710)
.60 and above	20	10
.50 - .59	27	26
.40 - .49	18	18
.39 and below	10	21
$\bar{X}$	.52	.47
s.d.	.10	.11

$\chi^2 = 7.26$
not significant at $p \leq .05$

the ATP Physics Test are a select group and that they have had about the same number of math courses as the male candidates. Many studies have reported differences in math achievement between males and females (e.g., Hilton and Berglund, 1974; Maccoby and Jacklin, 1974), but the more recent studies have reduced or eliminated the sex differences by controlling for relevant academic math experiences. In one such study using data from Project TALENT (Wise, 1979, p.9), it was found that "virtually all of the sex differences in twelfth grade math achievement could be explained in terms of sex differences in the number of math courses taken in high school." Further research on math background and how it relates to performance on specific types of test items in physics would help to clarify the results given in Table 9.

I-C. Are there items that discriminate differently for male and female candidates? If so, do these items have any discernable similarities? The r-biserial is a measure of the correlation of an item with the total test score. Values of the r-biserials should be significantly above zero, which they are in all cases, as shown in Table 8. If the test is working well for both males and females the r-biserial values for these two groups for each item should be similar. Jensen (1980) suggests ways of using the r-biserial correlations to study differences between two groups: the frequency distribution of the r-biserials for all the items should be similar for both groups; the number of nonsignificant r-biserials should be small and about the same for both groups; there should be a significant positive correlation between the r-biserials for the two groups; and, for any given item, the r-biserial correlations should not differ by more than twice the standard error of the difference between the two values.

The distributions of the r-biserial correlations for males and for females are shown in Table 10. A chi-square test, comparing these two distributions, indicates that they are not significantly different ($\chi^2 = 7.26$; not significant at $p \leq .05$). The mean r-biserial values differ by .05 and the standard deviations are about equal.

The lowest r-biserial value given in Table 8 is .19. Only four items have r-biserials below .30, one for males and three for females. Of the three below .30 for females, the r-biserial values for males ranged between .28 and .35. It appears that there is no difference between the males and females in terms of the number of nonsignificant r-biserial values that would help explain performance differences on this test.

The correlation between the r-biserials for male candidates and for female candidates

TABLE 11. Data on Speededness of the ATP Physics Test, December 1978

	Males (N = 802)	Females (N = 710)
Percent Completing the Full Test (answered item #75)	54.3%	48.0%
Percent Completing 75 Percent of the Test	98.9%	98.7%
Number of Items Reached by 80 Percent of the Candidates	71	71
Total Number of Items	75	75

is .70. This reflects a strong positive relationship between the two sets of r-biserial values.

The standard error of the difference for the r-biserials is .08. Following Jensen's (1980) suggestion that the r-biserials should not differ by more than twice the standard error of the difference, eight items were found to have differences in r-biserial correlations greater than .16, which indicates that these items might be discriminating differently for men and for women. Three of these eight items were identified as outliers (all relatively easier for males) in the analysis of delta values. Four were items on Mechanics, two on Electricity and Magnetism, and two on Optics and Waves. Six were Application items and two were Knowledge items. These eight items covered three item formats: Numerical, Pictorial, and Verbal. There is no clear item content or format pattern across these eight items that could help explain the differences in r-biserial correlations.

The use of r-biserials in examining test performance differences between groups has been minimal to date. Jensen (1980) points out some limitations of using r-biserial correlations in comparing performance of groups. These include relatively large sampling error levels, a limited range of values in the distribution of r-biserials, and, most importantly, the impact of item difficulty where the difficulty level is quite different for the two groups. Future studies on performance differences between groups should continue to explore the potential use of r-biserial correlations in identifying items that discriminate differently for two groups.

I-D. Are there similar patterns of performance with regard to speededness and omits for men and women? When performance differences are found on a timed test between two groups, one question which should be asked is whether the test is more speeded for one group than the other. Several variables can be considered in looking at speededness. They include the percentage of each group completing the full test, the percentage of each group completing 75 percent of the test, and the item reached by 80 percent of the candidates. Data on these three aspects of speededness are presented in Table 11. With the exception of the percentage of candidates completing the full test, the data in Table 11 are identical for both male and female candidates. The percent of candidates reaching each item is graphically shown in Figure 4.

The points on this graph confirm what can be seen in Table 11, that until the last four items there is no difference in the pattern of speededness for men and women. The variation near the end of the test, however, is not sufficient to explain the 10-point formula score performance difference on the test between the male and female candidates.

The next factor investigated in this study was the pattern of omits for men and women. In this analysis, an item was counted as an omit only if the candidate answered a

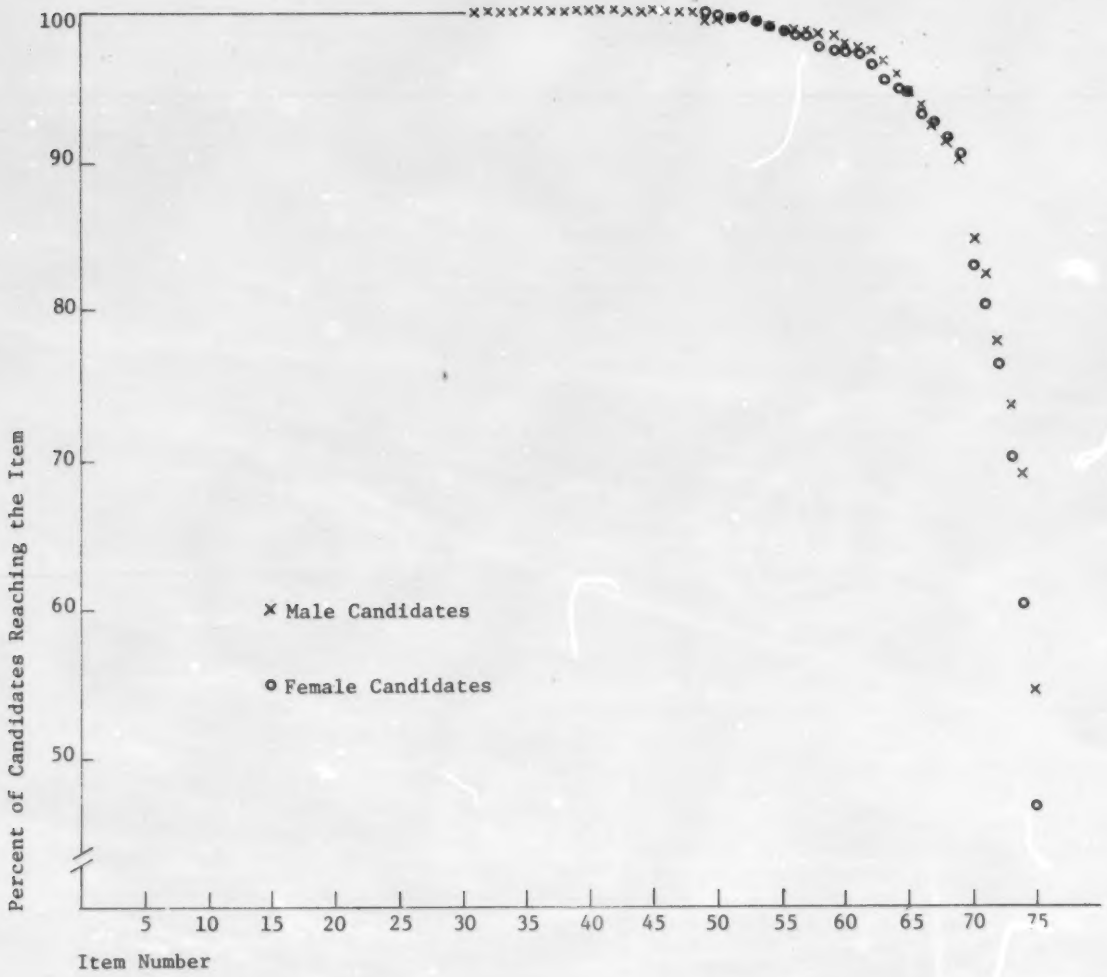

FIGURE 4. Patterns of Test Speededness for Male and Female Candidates, ATP Physics Test, December 1978

TABLE 12. Items Omitted by at Least Twice as Great a Proportion of Women as Men on the ATP Physics Test, December 1978 (N=802 Males; N=710 Females)

Item Number	Item Content	Bloom's Taxonomy	Item Format	Percent of Males Omitting	Percent of Females Omitting
2*	Mechanics	Application	Pictorial	3.7	10.2
3	Mechanics	Application	Pictorial	2.2	7.6
4	Mechanics	Application	Pictorial	1.7	5.7
5	Mechanics	Application	Pictorial	2.2	7.6
6	Mechanics	Application	Pictorial	4.8	11.0
7	Mechanics	Application	Pictorial	3.9	9.7
8	Optics and Waves	Knowledge	Verbal	2.9	6.3
9	Optics and Waves	Knowledge	Verbal	4.2	12.2
10	Optics and Waves	Application	Pictorial	3.9	12.3
11	Optics and Waves	Application	Pictorial	7.2	17.9
17	Mechanics	Application	Pictorial	3.1	8.0
18	Mechanics	Application	Pictorial	5.7	12.7
21	Mechanics	Knowledge	Verbal	2.5	5.9
25	Elec. and Magnetism	Comprehension/Applic.	Pictorial	2.4	5.5
32	Optics and Waves	Knowledge	Verbal	8.2	20.0
37	Mechanics	Application	Numerical	6.6	16.2
40*	Elec. and Magnetism	Application	Pictorial	3.2	7.4
41*	Elec. and Magnetism	Knowledge	Verbal	10.7	32.9
43*	Modern Physics	Knowl./Comp./Applic.	Verbal	16.9	34.5
44	Modern Physics	Application	Numerical	7.7	16.2
46	Heat and Kinetic Theory	Comprehension	Graphic	5.0	10.8
52*	Mechanics	Application	Numerical	10.3	26.5

* Outliers identified by analysis of delta values as the relatively easier items for males.

later item in the test, with the exception of the last item, in which case it was assumed that if the candidate reached item 74 he/she also reached item 75. The percentages of male and of female candidates who omitted each item are shown in Table 8. The means and standard deviations for these figures are provided in Table 6. As can be seen by these data, women are more apt to skip items than men. The correlation between the percent omit values for men and for women is .93, indicating that they have similar patterns for skipping items.

A review of the data in Table 8 shows that 22 items, almost one-third of the test, were omitted by twice as high a proportion of women as of men. They are listed in Table 12. Five of the seven outliers identified by the analysis of delta values as relatively easier items for males are on this list. These items also show the same pattern of item types that were found in the delta value outliers; that is, almost all the items omitted by twice as many women as men were Pictorial and Verbal items. There were two Numerical items (both of which contained Pictorial material) and one Graphic item. They were mostly Application and Knowledge questions and cut across all five content areas.

The data in Table 12 show something else of interest. The women were twice as apt to omit 10 of the first 11 items as men. These 11 items consist of three sets of related items (i.e., based on the same stimulus material) and most of these items have low delta values for women (i.e., easier items for women compared to other items in the test). This omit pattern on the early items in the test might indicate that some female candidates

are experiencing a higher level of test anxiety than male candidates. Previous research has indicated that women tend to score higher than men on test anxiety measures (Sarason, 1980). Omitting items early in the test might be a symptom of test anxiety.

In addition to looking at the ratio of the proportion of men and women omitting each item, the differences between the percent omit values for each item were also examined. The standard error of the difference for the percent omit distribution is 5.5. There were 16 items where the difference in percent omit for males and females was equal to or greater than twice the standard error of the difference, or 11.0. They are listed in Table 13. As in the other analyses in this study, these items cut across all five content areas. Most are Knowledge and Application items and in most cases (11 of 16) the format of the item is Verbal. Five of these 16 items were identified by the analysis of delta values as relatively easier for males.

The next step was to see if men and women were more apt to omit Verbal items than other types of items. For each item format (except Symbolic, for which there were only three items), the mean percent omit value was calculated for female candidates and for male candidates. The results are shown in Table 14. Female candidates were more apt to omit Verbal items than other types of items and least apt to omit Graphic items. Men were equally apt to omit Verbal and Numerical items. For all four types of items, women were more apt to omit the items than the men, although in the case of the Graphic items, of which there were only eight, the difference is small.

Another question concerning omits is how influential the level of item difficulty, as measured by the delta value for each item, is in candidates' omitting an item. The correlation between the percent omits and delta values for men was .63 and for women .61. Both groups tended, as would be expected, to omit the more difficult items. These same correlations were calculated for the first third of the test (items 1-25), the second third of the test (items 26-50), and the last third of the test (items 51-75). For both males and females the correlations for the first and second third of the test were in the range of .38 to .55, indicating a moderate-sized positive relationship between item difficulty and omitting items. However, for the last 25 items on the test, the correlation between delta values and percent omits was close to zero for both groups (.07 for men and .06 for women). Omitting items toward the end of the test might be more a factor of test speededness than of item difficulty. As shown in Figure 4, speededness was not a factor until the last third of the test.

The omit pattern is an area which should be investigated further. If female candidates are not willing to take risks and make an educated guess on as many items as male candidates, the result would be a lower total test performance level for women than men. Truly random guessing by all female candidates would have little, if any, impact on total scores since the test is formula scored (rights less one-quarter wrongs). However, educated guessing could increase their performance level.

I-E. Does an error analysis reveal a relationship between sex and choice of distractor? Although selection of distractors (wrong answers) does not affect total test scores, an investigation of differences in the selection of distractors might provide some insight into the processes that steer candidates away from the right answers. A distractor analysis was performed on the 14 items identified as outliers by the delta plots to see if the reason they were outliers was that one of the two groups was significantly more attracted to one of the distractors than the other group.

The distractor analysis was concerned with the distribution of responses to the four distractors for males and for females. A chi-square was computed for each of the 14 outliers, using a two (sex groups) by four (number of distractors per item) contingency table for each item. If either group were more attracted to a distractor than the other group, this could help to explain why fewer members of that group selected the right answer and thus had higher delta values.

The results of the distractor analysis, however, did not shed light on the perfor-

TABLE 13. Items Omitted by a Significantly Greater Percentage of Women than Men, ATP Physics Test, December 1978 (N=802 Males; N=710 Females)

Item Number	Item Content	Bloom's Taxonomy	Item Format	Percent of Males Omitting	Percent of Females Omitting
32	Optics and Waves	Knowledge	Verbal	8.2	20.0
41*	Elec. and Magnetism	Knowledge	Verbal	10.7	32.9
42*	Elec. and Magnetism	Application	Pictorial	31.4	57.8
43*	Modern Physics	Knowl./Comp./Applic.	Verbal	16.9	34.5
48	Optics and Waves	Knowledge	Verbal	11.6	22.9
51	Mechanics	Comp./Applic.	Symbolic	27.3	39.1
52*	Mechanics	Application	Numerical	10.3	26.5
54	Elec. and Magnetism	Knowledge/Comp.	Numerical	32.7	50.3
58	Modern Physics	Knowledge	Verbal	20.7	34.8
60	Modern Physics	Knowledge	Verbal	22.5	33.9
63*	Heat and Kinetic Theory	Knowledge/Applic.	Verbal	14.4	28.6
64	Optics and Waves	Knowledge	Verbal	24.3	39.6
65	Optics and Waves	Application	Numerical	17.1	30.2
66	Optics and Waves	Knowledge	Verbal	23.5	36.2
68	Mechanics	Knowl./Applic.	Verbal	21.4	38.0
71	Elec. and Magnetism	Application	Verbal	16.1	29.1

* Outliers identified by analysis of delta values as the relatively easier items for males.

TABLE 14. Average Percent Omit for Male and for Female Candidates by Item Format, ATP Physics Test, December 1978

Item Format	Males (N=802)	Females (N=710)	Ratio: F/M
Graphic (8 items)			
Mean	4.81	6.09	1.27
Standard Deviation	3.33	4.43	
Numerical (17 items)			
Mean	13.31	20.17	1.52
Standard Deviation	8.56	12.39	
Pictorial (24 items)			
Mean	8.36	15.40	1.84
Standard Deviation	7.61	12.00	
Verbal (27 items)			
Mean	13.66	23.17	1.70
Standard Deviation	7.06	10.72	

TABLE 15. Percent Choosing Distractors for Items Where Males and Females Differ in Their Most Popular Distractor, ATP Physics Test, December 1978 (N=802 Males; N=710 Females)

Item		Percent Choosing First Choice of Males	Percent Choosing First Choice of Females
2	Males	44.4%	31.6%
	Females	31.9%	35.7%
7	Males	53.9%	34.4%
	Females	40.3%	44.1%
8	Males	36.8%	34.2%
	Females	23.7%	37.3%
20	Males	51.8%	22.0%
	Females	30.7%	47.1%
25	Males	42.6%	41.0%
	Females	37.1%	40.5%
36	Males	54.4%	34.6%
	Females	39.6%	41.6%
41	Males	53.9%	19.4%
	Females	27.8%	42.6%
54	Males	38.4%	20.5%
	Females	27.1%	39.8%
60	Males	36.5%	36.1%
	Females	34.6%	35.6%
62	Males	39.4%	27.4%
	Females	28.9%	33.5%
70	Males	38.4%	31.9%
	Females	30.3%	39.9%
72	Males	37.6%	27.9%
	Females	27.9%	31.2%

mance differences. In 13 of the 14 cases, there was no significant difference in the distribution of males and of females selecting the distractors. In only one case was the difference significant. However, in this case male candidates were more apt to pick one distractor and female candidates were more apt to pick another distractor. About the same number of male and female candidates selected an incorrect distractor on this item.

The distractor analysis of outliers did not help explain the performance differences. However, a further exploratory analysis was done on all items in the test to investigate this possibility more completely. For each item, the four distractors were ranked by the number of male candidates who picked them and by the number of female candidates who picked them. Twelve items were found where the distractor picked most often by the male candidates was not the same as that picked most often by the female candidates. These items are listed in Table 15 along with the percent of male candidates who marked the wrong answer selected most by men and the one selected most by women, and the percent of female candidates who marked these two wrong answers. Some of the differences in selection of distractor are quite small (e.g., items 25 and 60), whereas others are larger (e.g., items 20 and 41).

The next step was, for each item, to look at the distractor that was the first choice of males and the distractor that was the first choice of females and then to identify the type of error they might be making. Without knowledge of the male/female differences in choice of response, the project's physics consultant classified the errors into four types: (1) misreading the question and possible answers; (2) making the situation more

complicated or complex than it actually is; (3) not understanding basic concepts, definitions, and processes; and (4) basing the response on similar situations that have different results.

In 10 of the 12 cases, female candidates tended to pick a distractor that indicated they did not understand basic concepts. In the other two, they appear to have misread the item. Male candidates appear to have made errors that can occur if one "knows too much"; in four cases they seem to have made the situation more complicated than it was, and in two others, to have based the answer on related experiences they might have had with similar situations that have different outcomes. On two items, the male candidates appear to have misread the item and on the remaining four to not understand basic concepts. These findings tend to support the popularly held notion that boys have more experiences, both academic and nonacademic, that are related to physics than do girls.

Data Analyses--Part II

The second part of this exploratory study focuses on characteristics of the candidates, such as total test performance and related academic background, rather than the test. The analyses for Part II were designed to provide information related to the following questions:

A. Do male and female candidates who achieve the same level of total test performance arrive at their scores in similar ways, e.g., do they have similar patterns of subtest score performance?

B. Does controlling for the number of physics courses or math courses reduce the male/female performance discrepancy on the total test or on subsections of the test?

The analyses on Part II were done on all candidates who took the December 1978 ATP Physics Test. However, when considering the impact of math background, it was only possible to consider those students for whom Student Descriptive Questionnaire (SDQ) data were in the data file.

Procedures for Creating Achievement Bands. In Part II-A, performance patterns for males and females at various levels of total ATP Physics Test performance were examined. Ordinarily, one would use an external criterion for separating candidates into "bands." In the absence of an external criterion, the total physics test score was used to create achievement intervals or bands. These bands were then used to investigate relative subscore performance.

Selecting the lower and upper limits for the achievement intervals served to reemphasize the differences between the distributions of scores for males and females. For example, if the top 15 percent of the candidates were selected (the formula score range of 55-75), there would be 677 males (17 percent of the males) and only 35 females (5 percent of the females). To use typical quintile or quartile bands would have created a situation in which the number of females in the top bands would be very small and consequently the results less reliable.

The alternative, which was used in this study, was to combine the score distribution for the females (N = 710) and the smaller sample of 802 males used for analysis in Part I. From this combined distribution, quintile interval score limits were established. In effect, the cutoff scores for the achievement band intervals were determined from a distribution in which the overall male performance patterns were maintained, but the sample sizes for males and females were roughly equal, rather than males outnumbering females five to one. This reduces the problem that would have resulted had the full sample been used to define the quintile intervals; that is, very few females would have appeared in the upper quintiles.

Although the reduced-sized male sample was used to define the achievement bands, analyses using these bands were performed on the entire sample of male candidates to get a

TABLE 16. Summary Statistics for Formula Scores by Achievement Band, ATP Physics Test, December 1978 (N = 4,015 Males; N = 710 Females)

Achievement Band	Formula Score Range	N		Median		Mean (Standard Deviation)	
		Males	Females	Males	Females	Males	Females
1. Lowest	-19 to 19	440	202	14.74	12.63	13.79 (4.07)	12.16 (4.44)
2.	20 to 27	605	174	22.93	22.54	23.00 (2.31)	22.62 (2.36)
3. Middle	28 to 36	799	154	31.31	30.85	31.29 (2.55)	30.82 (2.75)
4.	37 to 48	999	106	41.86	42.23	41.74 (3.49)	42.22 (3.51)
5. Highest	49 to 75	1,172	74	56.46	54.50	57.22 (6.39)	54.81 (5.50)

more stable estimate of their performance. Table 16 provides the formula score ranges for each band, the number of male and of female candidates in each band, and the mean and median formula scores within each band. As could be expected, even within the bands, the males tended to outperform slightly the females.

Results--Part II

II-A. Do male and female candidates who achieve the same level of total test performance arrive at their scores in similar ways? Do the males and females in a particular achievement interval evidence similar patterns of subscore performance? Are the patterns evidenced in one band repeated in the other bands? To investigate these questions, the distributions of subscores (number right) for males and females for each of the intervals were obtained. Figures 5, 6, and 7 graphically summarize these distributions for the content areas, the Bloom categories, and the item formats, respectively. Only findings from the highest, middle, and lowest bands are illustrated. Subscore patterns in the second and fourth bands evidence similar patterns to those illustrated in the figures, but were not included in order to simplify the presentation. In these three figures the following data are provided for each item category by sex: the range of scores, i.e., number of correct items (vertical line); the interquartile range (vertical bar); and medians (horizontal line within each vertical bar). The box around each subcategory of items delineates the range of possible scores.

For all three figures, the male/female performance patterns within each band and the subscore patterns between bands are remarkably similar. Across all three classification systems (content areas, Bloom categories, and item formats) and all five bands, the difference in median performance between males and females in a particular band for each category is never greater than two items; usually it is equal to or less than a one-item difference.

Another point illustrated by the figures is a relatively consistent pattern of subscore performance across bands. No particular content area, item format, or Bloom level of cognitive complexity stands out as being relatively easier or relatively harder for lower scoring candidates than it is for higher scoring candidates. These findings suggest that a change in the test content formula would not serve to reduce the male/female performance difference (such as the reduction illustrated in math by Donlon, 1973).

II-B. Does controlling for the number of physics or mathematics courses reduce the male/female discrepancy on the total test or on subsections of the test?

Physics: All candidates for the ATP Physics Test were asked on the front of the test booklet to specify the number of semesters of physics that they had completed, including the current semester if it was more than half over. As was shown in Table 1, about 95

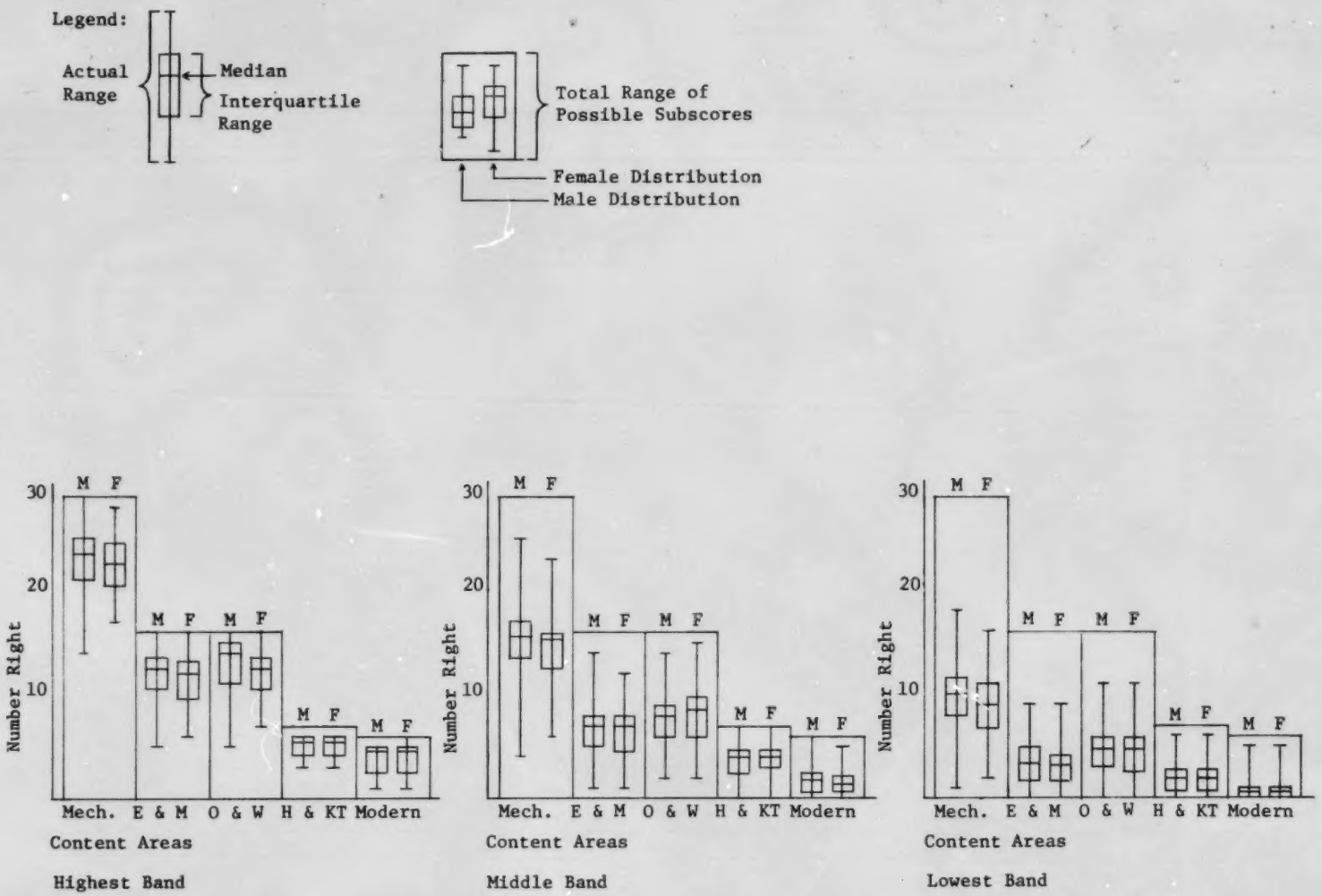

FIGURE 5. Distribution of Subscores by Content Area for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates

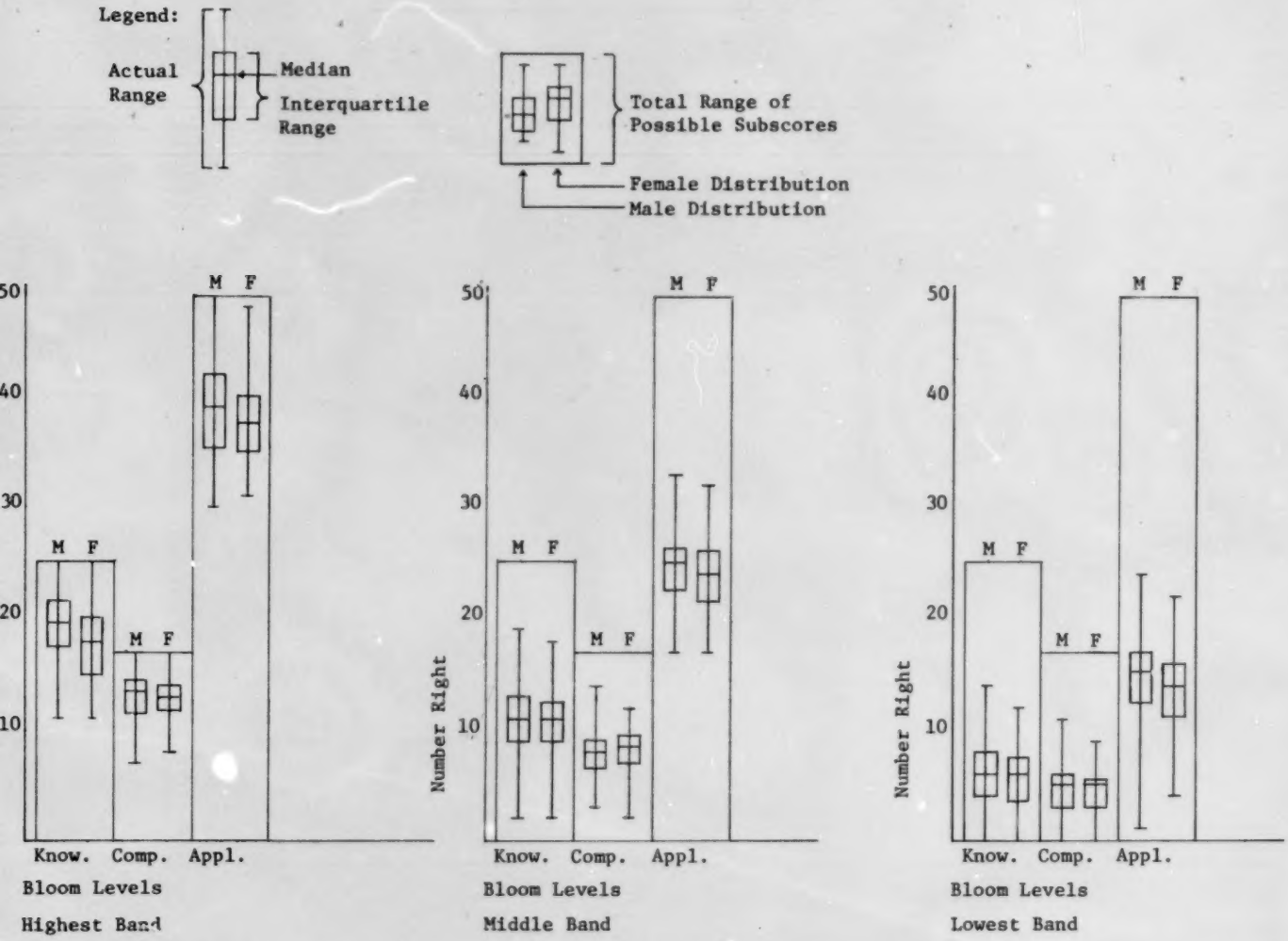

FIGURE 6. Distribution of Subscores by Bloom Categories of Cognitive Complexity for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates

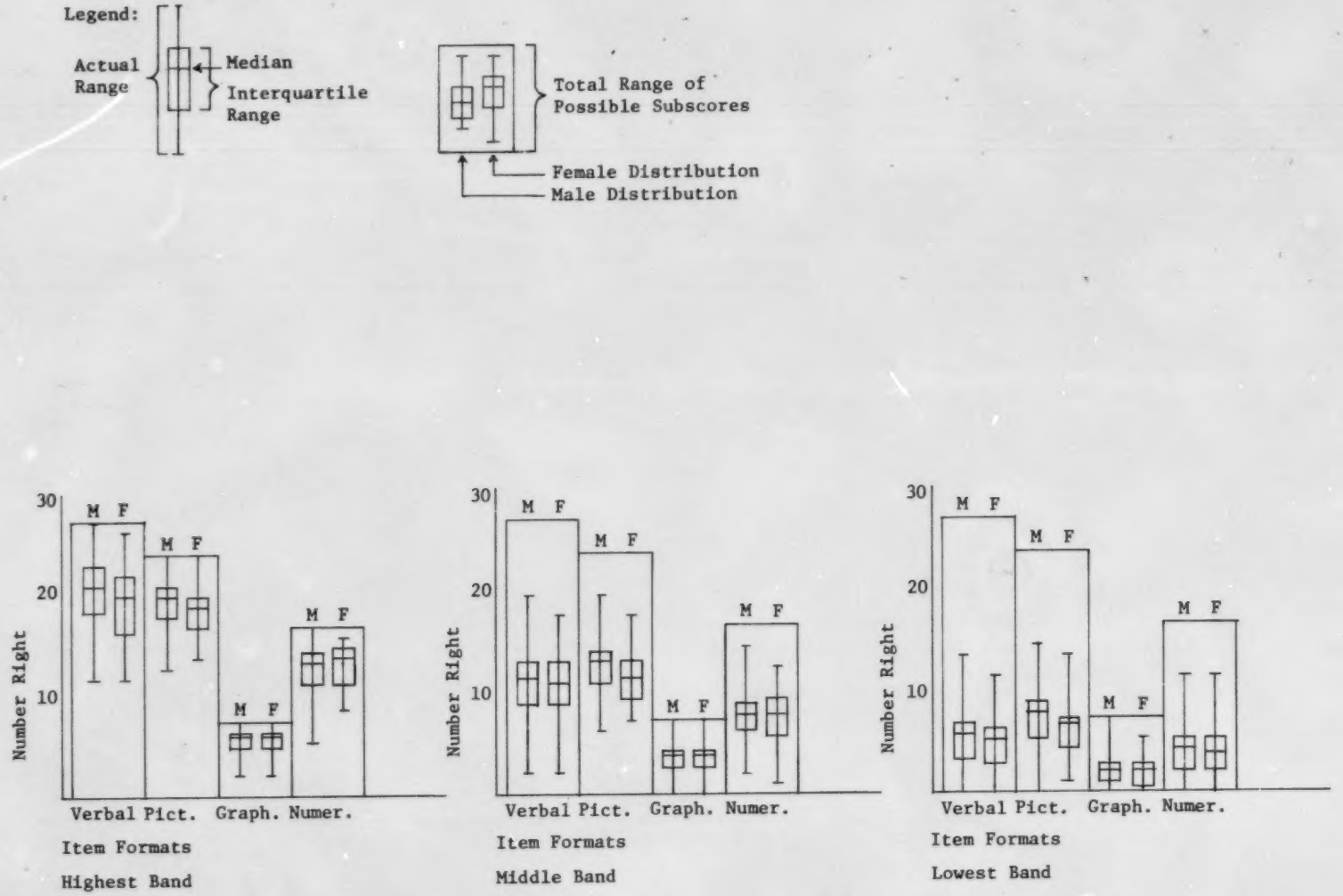

FIGURE 7. Distribution of Subscores by Item Format for Highest, Middle, and Lowest Achievement Bands for Male and Female ATP Physics Test Candidates

percent of the candidates were seniors. Because physics is a subject usually taken in the senior year, many of the candidates electing to take the ATP Physics Test in December were seniors who had completed only about half of their first physics course. This was true for about one-third of the candidates. For this reason, comparisons were made not just of the proportions of males and females by numbers of semesters of physics (see Table 2), but also of their relative performance by subscores and on the total ATP Physics Test by numbers of semesters of physics.

Figures 8, 9, and 10 provide the means for the total test formula score and the content, Bloom level, and item format subscores (number right) by number of semesters of physics for males and females. As expected, the mean performance of both males and females tends to increase in all categories as the number of semesters of physics increases. A curious exception to this trend is found on the total test (see Figure 9) and in most of the content areas for third- and fourth-semester males (see Figure 8). One could speculate that the decline in performance for fourth-semester male candidates could be related to a lapse in time between the completion of a high school course (usually spring) and the December test administration. However, this pattern does not hold true for women; i.e., mean female performance on the total test and on all subtests increases with more physics course work. Further investigation would be necessary to determine if this inconsistency is evidenced on other test dates or if SDQ data could provide some insight into this.

Presumably, some of the content areas represented on the test would not be covered during the first semester of physics, in which case related background experiences could be an important factor in performance differences of first-semester male and female students. This reliance on experience rather than on strictly academic learning might give males an advantage in one or more of the content areas.

However, as shown in Figure 8, no one content area stands out as evidence of a male advantage; rather, there is a consistent pattern of male superiority in test performance across all five content areas. The mean performance of first-semester male candidates is equal to or greater than the mean performance of both first- and second-semester female candidates in all content areas. The finding that this discrepancy does not exist for candidates reporting four semesters of physics suggests the plausibility of a number of explanations such as: (1) Males have an experiential advantage over females in physics that diminishes as women gain in physics background; (2) ATP physics candidates with more than two semesters of physics background are a unique subset of the population of physics candidates. Females who have elected to advance in this area far beyond minimum high school requirements may have experiences, attitudes, interests, aiblities, and/or self-confidence which make them more similar to the male candidates with three or four semesters of physics. Investigating the likelihood of these explanations is beyond the scope of this exploratory study.

As shown in Figures 8, 9, and 10, when the reported number of semesters of physics increases beyond two semesters, the discrepancy between male and female performance decreases. Across all subtests and on the test as a whole, the discrepancy between male and female performance for first-semester candidates is similar to the discrepancy for second-semester candidates. However, by the third semester the male/female gap begins to close and by the fourth semester, with one exception (Verbal Item Format Subscore; significant at $p \leq .05$), the mean performance for males and females is equal or not significantly different. This pattern does not appear to be due to a ceiling effect, since mean scores are well below the total possible score in each case.

One reservation about these findings is that there is a sharp decline in the numbers of males and females who take three or more semesters of physics as compared to those who take fewer. Although the proportions of male and female candidates with four or more semesters of physics are comparable, the lower participation rate of females on the test as a whole results in a small sample size for fourth-semester females ($N = 22$). Thus, it will be important to reaffirm these findings in future studies. Candidate information on the types of physics course (e.g., Advanced Placement physics course, physical science

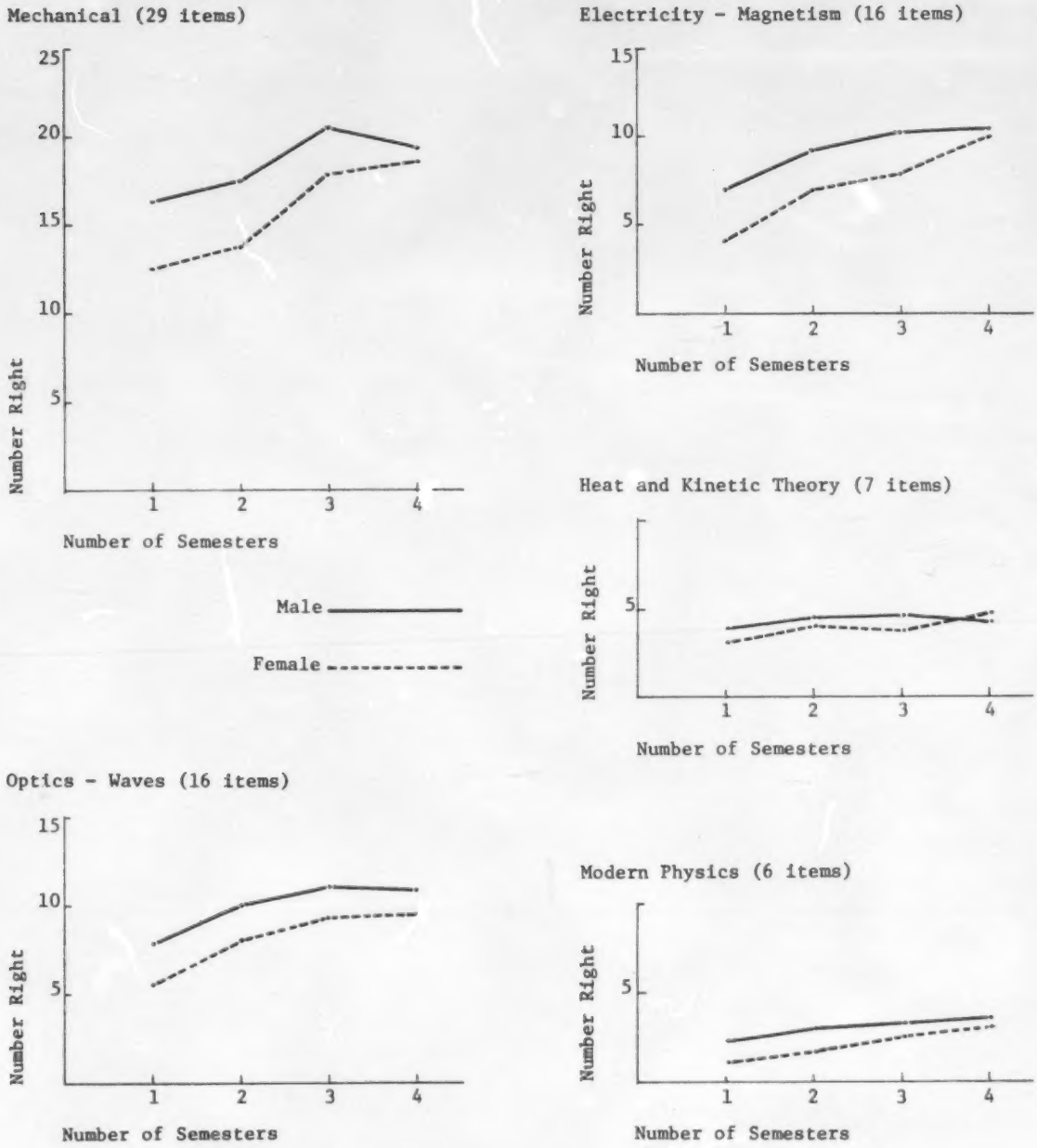

FIGURE 8. Mean Scores for Male and Female ATP Physics Test Candidates by Content Area by Number of Semesters of Physics

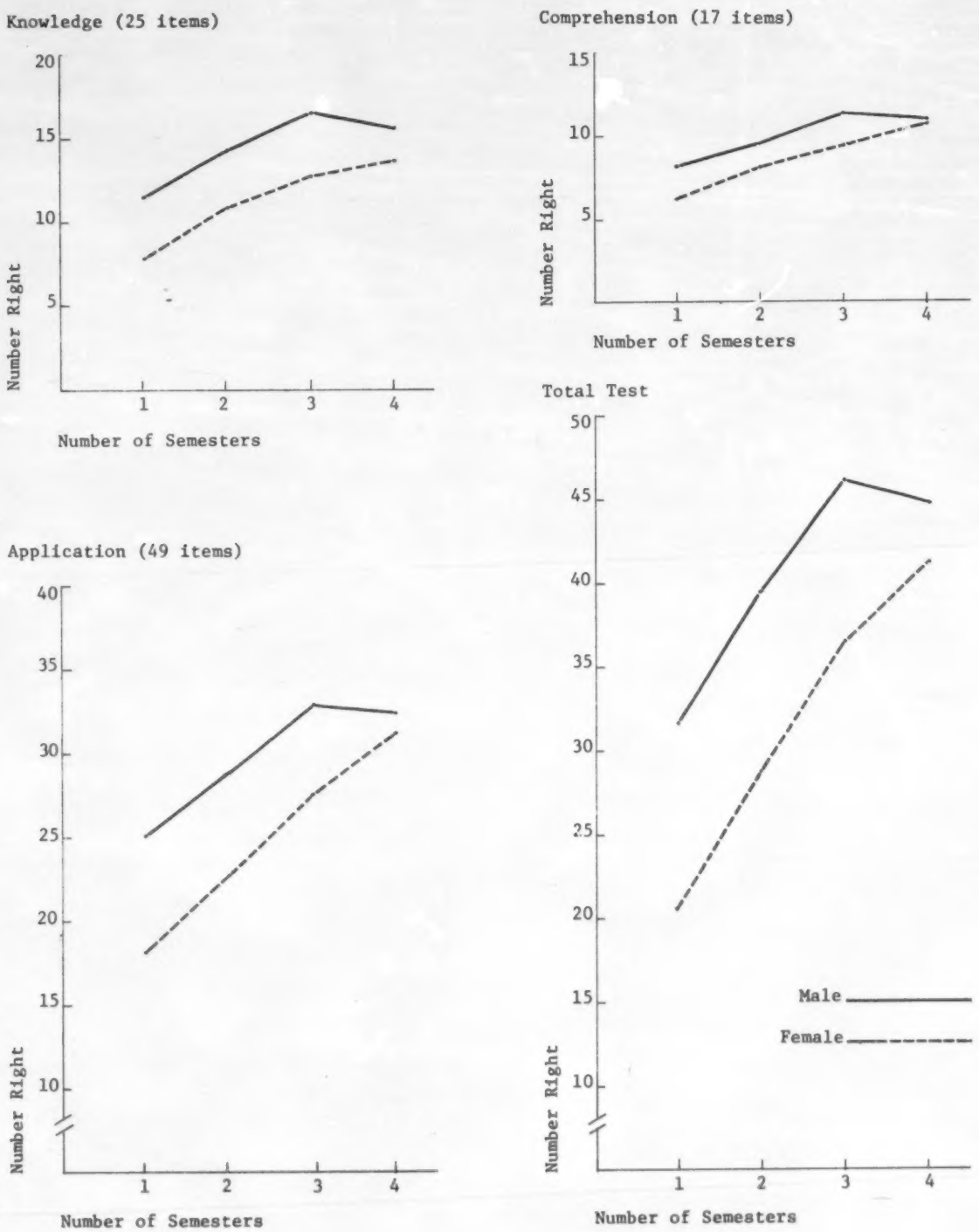

FIGURE 9. Mean Scores for Male and Female ATP Physics Test Candidates by Bloom Categories of Cognitive Complexity and on Total Test by Number of Semesters of Physics

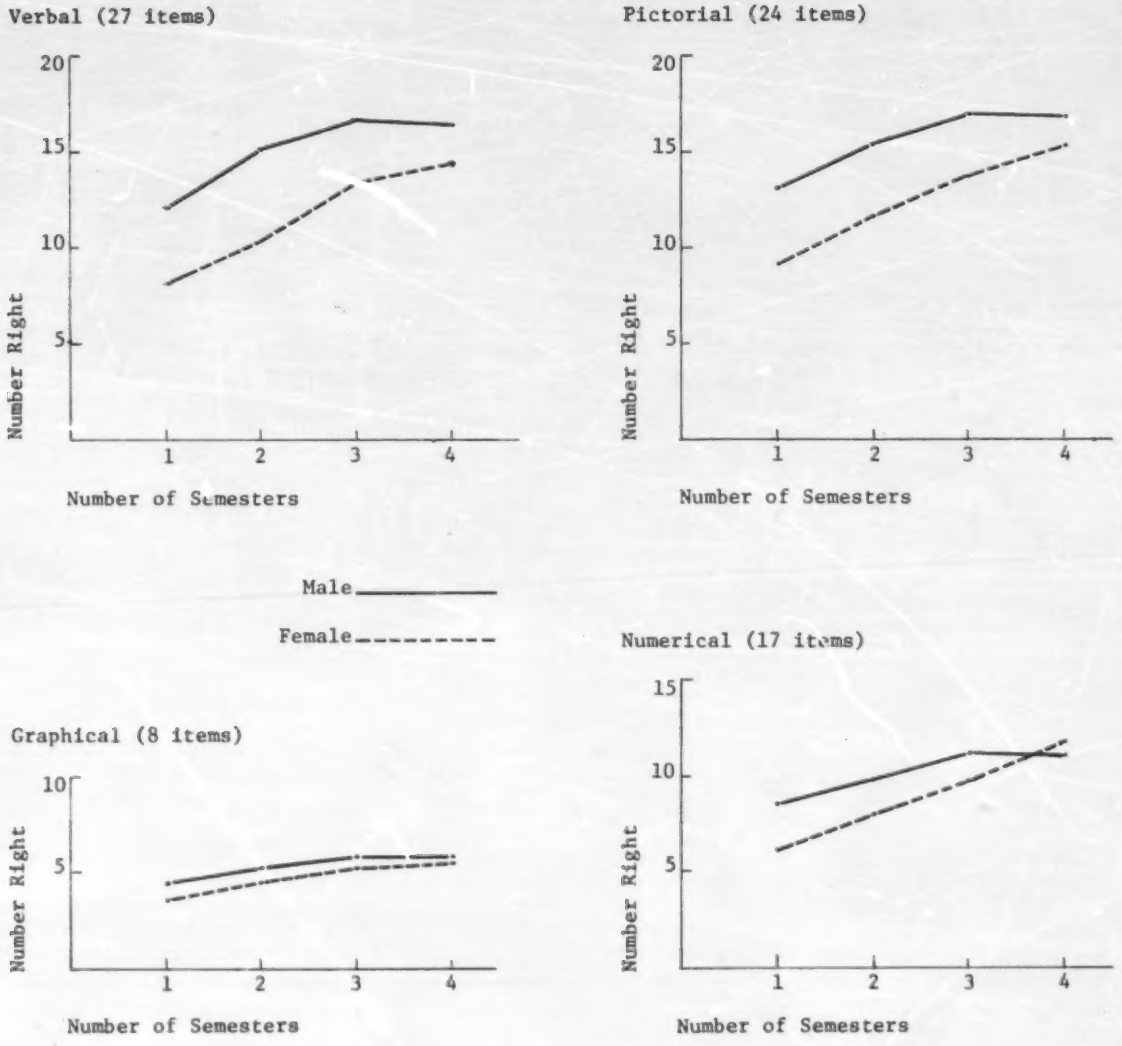

FIGURE 10. Mean Scores for Male and Female ATP Physics Test Candidates by Item Format by Number of Semesters of Physics

survey course) and the amount of physics lab work would be useful in future studies on the relationship between physics course work and test performance.

Mathematics: Success in the physical sciences is frequently linked to competency in math. Generally a prerequisite for high school physics is at least one course in algebra. Further course work in geometry and trigonometry, while not essential to learning physics, can be of great benefit in learning about some topics in physics (e.g., optics, sound waves, vectors). Therefore, test performance by level of math background was also examined in this study.

As was mentioned earlier in this report, the source of math background data was the Student Descriptive Questionnaire. More than 96 percent of the candidates providing this information were seniors. Since the question asks how many years of math the candidate expects to complete before high school graduation, collapsing data across grades is not reasonable, and sample sizes in other grades were too small to analyze separately.

Figure 11 presents the means of the total test performance by sex for various levels of math background. It shows that for both sexes, as expected, the number of years of math background is positively related to performance on the total test. It is also positively related to both the quantitative and qualitative subscores; the more years of high school math a candidate has, the higher the score on both the quantitative and the qualitative items.

On all three scores (Total, Quantitative, Qualitative) the mean formula scores of women with more than four years of math are lower than the corresponding mean formula scores for men with three years of math. The females' mean scores are within a point of the corresponding means for men reporting two years of math or less. This finding seems to indicate that the performance advantage of men over women in physics is independent of the self-report of the number of years of high school math. It is important to note, however, that it was not possible to compare the type or level of math courses which males and females have taken or their level of math achievement without obtaining additional data. Further investigation is needed in this area.

SUMMARY

The purpose of this exploratory investigation was to examine the differences in performance between males and females on the ATP Physics Test in an effort to gain insights into the sources of these differences. This was done by first looking at factors in the test that might help to account for these differences and then by relating test performance to other factors that might provide insights into possible source(s) of the differences.

To examine factors in the test that could help account for the sex differences, a standard item analysis was performed separately by sex on the entire sample of female candidates (N = 710) and a similar-sized random sample of the male candidates (N = 802) who took the test in December 1978. Using a formula score that incorporates a correction factor for guessing (number of items right, minus one-quarter the number wrong), the men performed about 10 points higher than the women. The mean formula score for males was 38.19 and for females 27.99 on the 75-item test.

Test summary statistics (internal consistency reliability, standard error of measurement, r-biserials) were similar for men and women. However, the score distribution for women tended to be more positively skewed than the distribution for men, the mean item difficulty value (delta) for women was higher than that of the men, and the mean percent omit level was higher for women than men. These differences, along with the difference in mean performance, all support the finding that the test is more difficult for the female candidates than the male candidates.

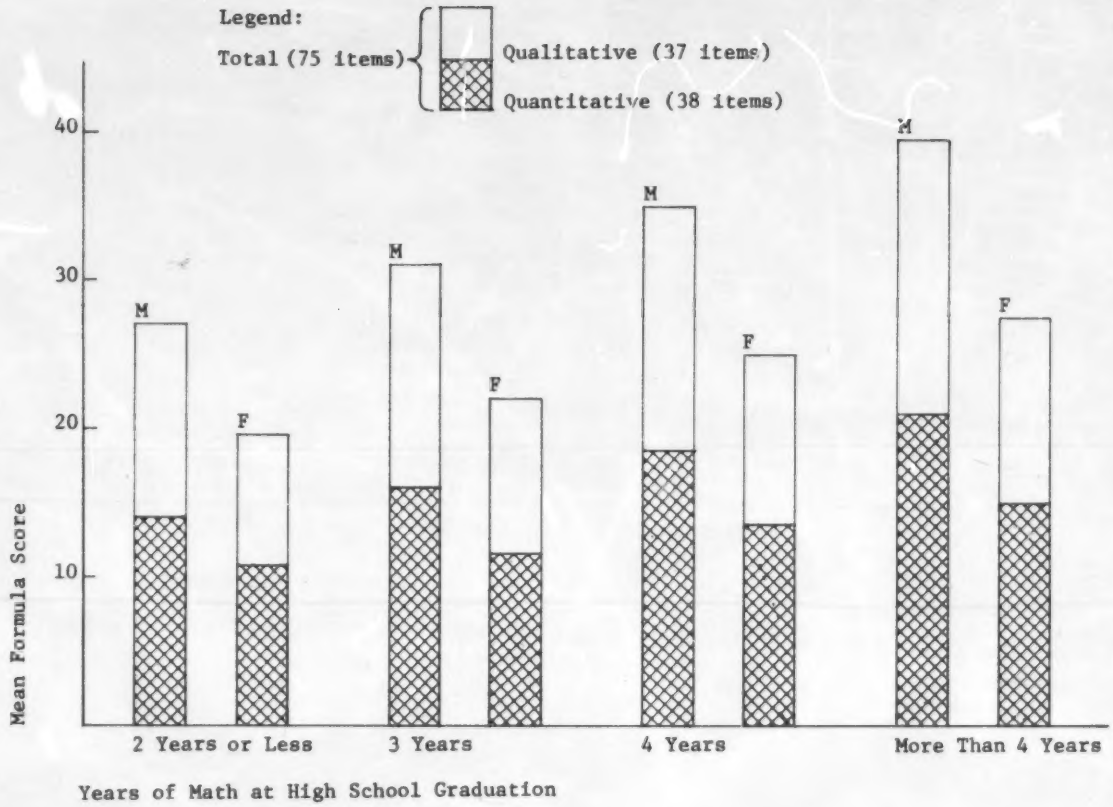

FIGURE 11. Means of Total Formula Scores and Quantitative and Qualitative Subscores by Number of Years of Math at High School Graduation by Sex for Grade 12 ATP Physics Candidates

To investigate whether there were differential patterns of item difficulty, the delta values for men and women for each item were cross plotted and outliers (i.e., items that are relatively easier items for men and items that are relatively easier for women) were identified. The seven outliers that were relatively easier items for men and the seven that were relatively easier items for women cut across all five content areas of the test. Of the three areas of cognitive processing, items favoring men tended to be classified as Application and Knowledge whereas those favoring women tended to be classified as Application and Comprehension items. This is not consistent with the earlier findings of Carlson (1976) which point to Physics Understanding/Comprehension items as the area of greatest male/female difference on the IEA Science Tests with males consistently outperforming the females from as early as grade three.

Another notable inconsistency with past research pertained to item formats. Contrary to what might be expected, the results showed that outliers that tended to be easier items for males were Pictorial and Verbal, while those that were easier items for females, relative to other items, were Graphic, Numerical, and Symbolic items.

Although there were no major differences in test speededness data for male and female candidates, women were much more apt than men to omit items. Male candidates omitted 11 percent of the items they reached whereas female candidates omitted almost 19 percent of the items they reached. On about one-third of the items (22 of 75), at least twice as high a proportion of women omitted the item as men. Ten of the first 11 items were among these 22; this may indicate that the female candidates were experiencing a higher level of test anxiety than the male candidates. It is also possible that women may be less willing to reason and to take risks during the test and consequently are less willing to make educated guesses. The tendencies for females to avoid reasoning and to be reluctant to take risks have been found in other situations as well (Linn, 1980; Maccoby and Jacklin, 1974; Sarason, 1980). Further investigation is indicated to determine what the effect on performance would be if women were encouraged to make educated guesses.

In the second part of the investigation the focus of inquiry was redirected from the test to factors related to the candidate that, when combined with subtest or total test performance, might help identify most of the sources of the performance differences.

One line of inquiry was to look at the patterns of subscore performance for candidates at particular levels or ranges of overall achievement on the ATP Physics Test in order to determine if men and women who achieved the same total performance level on the test arrived at this level in similar ways. This made it possible to look at male/female patterns within a range of achievement and also to compare the patterns of subscore performance across levels of achievement. Subscores that were investigated by achievement levels were content areas, item formats, and Bloom classifications. Five intervals or achievement bands were used in this analysis.

For all three types of classifications, the male/female performance patterns within each performance achievement interval and subscore patterns between achievement intervals were remarkably similar. At none of the five levels of achievement did the males evidence a particular advantage over females, or vice versa, as measured by the three classification systems. This suggests that a change in the test content formula or distributions of item types would not serve to reduce the male/female performance differences.

Academic background was also considered as a factor that might help explain the sex differences. ATP Physics Test candidates were asked to indicate the number of semesters of physics that they had completed. The expected number of years of high school mathematics to be completed by high school graduation was a question on the Student Descriptive Questionnaire (an optional series of questions about candidate interests, experiences, activities, and plans) that about 30 percent of the candidates completed as part of the December 1978 administration.

The number of semesters of physics that a test candidate has completed proved to be an important variable in interpreting some of the sex differences on the ATP Physics

Test. It was found that as the reported number of semesters of physics increased beyond two semesters, the discrepancy between male/female performance diminished to a nonsignificant level. This was true with one exception (the "Verbal" item format subscores) for all subscore categories. Although caution should be used in generalizing about this finding because of the small sample size, exposure to physics course work might be a critical variable in reducing the male/female gap in performance.

Number of years of mathematics was related to overall performance and to performance on both quantitative and qualitative items. However, sex differences remained unaltered regardless of the number of years of math. Further research is needed on these and other candidate characteristics to help identify factors external to the test that are related to the performance differences in physics between men and women.

Although no simple explanation for the performance discrepancy was found, it was possible to identify some factors that were related to sex differences in test performance. Some of the findings related to performance differences in physics failed to support the explanations given in prior research. The major findings are summarized here:

1. When the test is analyzed separately by sex, the test summary statistics and item analysis show that the ATP Physics Test is a reliable instrument for both men and women.

2. There were only three items of the 75 on the test for which the proportion of women passing the item equaled or exceeded the proportion of men passing the item.

3. By comparing within-sex-group patterns of item difficulty, it was possible to determine which of the items men found to be relatively easier and which of the items women found to be relatively easier compared to other items. The following conclusions were derived from this analysis:

   A. None of the five content areas represented on the test (Mechanics, Electricity and Magnetism, Optics and Waves, Heat and Kinetic Theory, Modern Physics) stood out as being relatively easier for males or for females, compared to the other areas.

   B. Items were classified into three levels of cognitive complexity according to Bloom's taxonomy: Knowledge, Comprehension, and Application. Men tended to find the Application and Knowledge items easier than the Comprehension items. Women found the Application and Comprehension items easier than the Knowledge items.

   C. Items were categorized into five item formats: Graphic, Numerical, Pictorial, Symbolic, and Verbal. Contrary to previous research findings, items that were relatively easier for men were Verbal and Pictorial. Women, on the other hand, found Graphic, Numerical, and Symbolic items to be relatively easier than the Verbal and Pictorial items.

4. Almost one-third of the items on the test were omitted by twice as high a proportion of women as men. The types of items that women tended to omit were the same types of items identified as the relatively easier items for men: Pictorial, Verbal, Application, and Knowledge.

5. Across all content area subtests and on the test as a whole, the male/female performance discrepancy for first-semester candidates was similar to the discrepancy for second-semester candidates. As the reported number of semesters of physics increased beyond two semesters, the gap between male/female performance narrowed. The only significant difference in performance for fourth-semester male and female physics candidates was on Verbal items, where men surpassed women.

6. The performance advantage of men over women is independent of the self-reported number of years of high school math. No information was available on the types of mathematics courses taken.

The results of this study point to several areas in which further investigation is needed. With regard to the relationships between item types and performance, the conclusions stated above need to be replicated on other forms of the test, thus increasing the item pools in each classification. The need for further investigation into test-taking behavior and its impact (as evidenced by sex differences in omit patterns) is also indicated.

Lastly, the consistent sex differences in performance evidenced in this test (which only seem to diminish as semesters of physics background increased beyond what is normally expected of college-bound youth) amplify the need to look beyond the test and to increase our understanding of the factors that influence and enhance science participation and involvement.

REFERENCES

Aiken, Lewis R. "Update on Attitudes and Other Affective Variables in Learning Mathematics," Review of Educational Research, 46:2 (1976): 293-311.

Anastasi, Anne. Differential Psychology. New York: The Macmillan Company, 1958.

Andersen, Charles, Ed. A Fact Book on Higher Education. 4th issue. Washington, D.C.: American Council on Education, 1976.

Angoff, William H., and Ford, Susan F. "Item-Race Interaction on a Test of Scholastic Aptitude," Journal of Educational Measurement, 10:2 (1973): 95-106.

Baker, Sheldon R., and Talley, Lawrence H. "The Relationship of Visualization Skills to Achievement in Freshman Chemistry," Journal of Chemical Education, 49 (1972): 775-776.

Baker, Sheldon R., and Talley, Lawrence H. "Visualization Skills as a Component of Aptitude for Chemistry--A Construct Validation Study," Journal of Research in Science Teaching, 11:2 (1974): 95-97.

Bloom, Benjamin S. Taxonomy of Educational Objectives. Handbook I: Cognitive Domain. New York: David McKay Co., 1956.

Bock, R. Darrell, and Kolakowski, Donald. "Further Evidence of Sex-Linked Major-Gene Influence on Human Spatial Visualizing Ability," American Journal of Human Genetics, 25 (1973): 1-14.

Carlson, Marthena. "Sex Differences in Performance on the IEA Science Test Items." Ph.D. dissertation, University of Pittsburgh, 1976. (University Microfilms International #77-21, 262)

Casserly, Patricia Lund. "What Can Schools Do to Enhance Participation and Achievement of Young Women in Mathematics and the Physical Sciences?" Paper presented at the American Association for the Advancement of Science, February 16, 1978.

Centra, John A. Women, Men and the Doctorate. Princeton, N.J.: Educational Testing Service, 1974.

Coffman, William E. "Sex Differences in Response to Items in an Aptitude Test." In Eighteenth Yearbook, National Council on Measurement in Education, pp. 117-124. Washington, D.C.: National Council on Measurement in Education, 1961.

College Board. The Physics Examinations of the College Board. New York: College Entrance Examination Board, 1978.

Comber, L. C., and Keeves, John P. Science Education in Nineteen Countries: An Empirical Study. New York: John Wiley & Sons, 1973.

de Wolf, Virginia A. "High School Mathematics Preparation and Sex Differences in Quantitative Abilities," Draft of paper (1980) to be published in Psychology of Women Quarterly.

Donlon, Thomas F. Content Factors in Sex Differences on Test Questions. (Research Memorandum 73-28) Princeton, N.J.: Educational Testing Service, 1973.

Donlon, Thomas F.; Hicks, Marilyn M., and Wallmark, Madeline M. "Sex Differences in Item Responses on the Graduate Record Examination," Applied Psychological Measurement, 4:1 (1980): 9-20.

Fennema, Elizabeth, and Sherman, Julia. "Sex-Related Differences in Mathematics Achievement, Spatial Visualization and Affective Factors," American Educational Research Journal, 14:1 (1977): 51-71.

Hartnett, Rodney T., and Feldmesser, Robert A. "College Admissions Testing and the Myth of Selectivity: Unresolved Questions and Needed Research," American Association of Higher Education Bulletin, 32:7 (1980): 3-6.

Hilton, Thomas L., and Berglund, Gosta W. "Sex Differences in Mathematics Achievement--A Longitudinal Study," The Journal of Educational Research, 67:5 (1974): 231-237.

Hoffman, L. R., and Maier, N. R. F. "Social Factors Influencing Problem-Solving in Women," Journal of Personality and Social Psychology, 4 (1966): 382-390.

Husen, Torsten; Fägerlind, Ingemar; and Liljefors, Robert. "Sex Differences in Science Achievement and Attitudes: A Swedish Analysis by Grade Level," Comparative Education Review, 18 (1974): 292-304.

Jensen, Arthur R. Bias in Mental Testing. New York: The Free Press, 1980.

Keeves, John P. "Differences between the Sexes in Mathematics and Science Courses," International Review of Education, 19 (1973): 47-62.

Kelly, Alison. Girls and Science: An International Study of Sex Differences in School Science Achievement. (International Association for the Evaluation of Educational Achievement, Monograph No. 9) Stockholm: Almquist and Wiksell, 1978.

Kelly, Alison. "Women in Science: A Bibliographic Review," Durham Research Review, (1976): 1092-1108.

Linn, Marcia C. Changing the Filter on Adolescent Cognitive Behavior: A Consideration of Context Effects. Berkeley, Calif.: University of California, Lawrence Hall of Science, 1980.

Lord, Frederic M. Practical Applications of Item Characteristic Curve Theory. (Research Bulletin 77-3) Princeton, N.J.: Educational Testing Service, 1977.

Maccoby, Eleanor E., and Jacklin, Carol N. The Psychology of Sex Differences. Stanford, Calif.: Stanford University Press, 1974.

Max, Claire Ellen. "Opportunities for Women in Physics." Paper presented at the American Physical Society, January 23, 1978, San Francisco.

Milton, G. A. "Sex Differences in Problem Solving as a Function of Role Appropriateness of the Problem Content," Psychological Reports, 5 (1959): 705-708.

National Assessment of Educational Progress. Changes in Science Performance, 1969-1973; Exercise Volume. (Science Report No. 04-5-20) Washington, D.C.: Government Printing Office, 1975.

Pfeiffenberger, George W. "National Testing in Physics: Recent Trends and Future Implications," The Physics Teacher, 12:7 (1974): 391-395.

Rossi, Alice S. "Women in Science: Why So Few?" Science, 148 (1965): 1196-1202.

Sarason, Irwin G., Ed. Test Anxiety: Theory, Research, and Applications. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Scheuneman, Janice. "A New Method of Assessing Bias in Test Items." Paper presented at the annual meeting of the American Educational Research Association, April 1975, at Washington, D.C.

Sells, Lucy W. "The Mathematics Filter and the Education of Women and Minorities." Paper presented at the annual meeting of the American Association for the Advancement of Science, January 1976.

Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co., 1956.

Stafford, Richard E. "Sex Differences in Spatial Visualization as Evidence of Sex-Linked Inheritance," Perceptual and Motor Skills, 13 (1961): 428.

Stake, Jayne E., and Granger, Charles R. "Same-Sex and Opposite-Sex Teacher Model Influences on Science Career Commitment Among High School Students," Journal of Educational Psychology, 70:2 (1978): 180-186.

Stanley, Julian C.; Keating, Daniel P.; and Fox, Lynn H., Eds. Mathematical Talent: Discovery, Description, and Development. Baltimore: Johns Hopkins University Press, 1974.

Stern, June. College Board Item Bias Study of the Scholastic Aptitude Test and the Test of Standard Written English. (SR 78-56) Princeton, N.J.: Educational Testing Service, 1978.

Strassberg-Rosenberg, B., and Donlon, Thomas F. "Content Influences on Sex Differences in Performance on Aptitude Tests." Paper presented at the annual meeting of the National Council on Measurement in Education, April 1975, at Washington, D.C.

Walberg, Herbert J. "Physics, Femininity and Creativity," Developmental Psychology, 1:1 (1969): 47-54.

Walker, Helen M., and Lev, Joseph. Statistical Inference. New York: Henry Holt and Company, 1954.

Wild, Cheryl L., and Durso, Robin. Effect of Increased Test-Taking Time on Test Scores by Ethnic Group, Age, and Sex. (Graduate Record Examinations Board Research Reports, GREB No. 76-6R) Princeton, N.J.: Educational Testing Service, 1979.

Wise, Lauress L. Long-Term Consequences of Sex Differences in High School Mathematics Education. Paper presented at the annual meeting of the American Educational Research Association, April 8, 1979, at San Francisco.

Wright, B. D.; Mead, R. J.; and Draba, R. E. Detecting and Correcting Test Item Bias with a Logistic Response Model. (RM No. 22) Chicago: University of Chicago, Department of Education, 1976.